UNIVERSITY OF CALIFORNIA

Santa Barbara

Development of bioinformatic tools to identify and characterize linear protein epitopes

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Chemical Engineering

by

Michael Louis Paull

Committee in charge:

Professor Patrick S. Daugherty, Chair

Professor M. Scott Shell

Professor Michelle A. O'Malley

Professor Ambuj K. Singh

Professor Dennis O. Clegg

June 2018

The dissertation of Michael Louis Paull is approved.

_____

M. Scott Shell

_____

Michelle A. O'Malley

_____

Ambuj K. Singh

_____

Dennis O. Clegg

_____

Patrick S. Daugherty, Committee Chair

May 2018

Development of bioinformatic tools to identify and characterize linear protein epitopes

ACKNOWLEDGEMENTS

"Let your house be a meeting place for sages; and sit in the dust of their feet; and drink in their words thirstily" – Pirkei Avos 1:4

I would like to thank UCSB, the department of chemical engineering, and everyone else who let me participate in such stimulating research. Thank you to my advisor Patrick Daugherty who gave me the freedom and resources to follow my imagination. You taught me the importance of always having a high-level sense of the direction of my research. I would like to thank the rest of my committee, Michelle O'Malley, Scott Shell, Dennis Clegg, and Ambuj Singh for encouraging me along this journey and providing useful advice. Thank you to the Daugherty and Plaxco groups for fun get-togethers and daily commiseration. A special thanks to Joel who been a remarkable colleague and friend. I would like to thank my friends in chemical engineering for trying new restaurants with me and keeping this process enjoyable.

I would like to thank the Jewish community and all the opportunities to learn and travel that they have given me. I especially would like to thank Rabbis Hanfling and Kaufman for giving me perspective, shaping my future, and being good friends. Thank you to all my friends from the East Coast. I have really enjoyed periodically chatting and playing video games with you. Thank you to all my family who has been extremely supportive throughout this whole process. To my grandmothers, Maddy and Mama, thank you for instilling in me the value of education and family. To my grandfathers, Papa and Richard, thank you for setting the foundation for me to be an engineer, and may your memory be a blessing. To my sister, I've really enjoyed going down to see you in LA throughout my PhD.

To my brother, I look forward to all the Dr. Paull jokes that we will make. To my mom and dad, I owe so much to both of you. You have been unwaveringly supportive of every life decision I've made. Our weekly conversations have always allayed my concerns and given me faith that I could accomplish whatever I put my mind to. To everyone who helped me along the way, I offer my sincerest thanks.

VITA OF MICHAEL LOUIS PAULL

June 2018

EDUCATION

2013-2018 Doctor of Philosophy in Chemical Engineering
University of California, Santa Barbara

2009-2013 Bachelor of Science in Chemical and Biomolecular Engineering
Graduated Magna Cum Laude
Cornell University

PROFESSIONAL EXPERIENCE

2013-2018 Graduate Student Researcher and Teaching Assistant
Chemical Engineering, University of California, Santa Barbara
Research Advisor: Patrick S. Daugherty

2011-2013 Undergraduate Student Researcher and Teaching Assistant
Chemical and Biomolecular Engineering, Cornell University
Research Advisor: Jeffrey D. Varner

PUBLICATIONS AND PATENTS

Paull ML, Daugherty PS: Mapping serum antibody repertoires using peptide libraries. *Current Opinion in Chemical Engineering* 19 (2018): 21-26.

Paull ML, Johnston T, Ibsen KN, Bozekowski JD, Daugherty PS: A general approach for identifying protein epitopes targeted by antibody repertoires using whole proteomes. *In Preparation.*

Paull ML, Bozekowski JD, Daugherty PS: Mapping antibody binding using multiplexed epitope substitution analysis. *In Preparation*.

PRESENTATIONS

- "Serum Antibody Target Discovery Using Pattern Tiling"
  255th ACS National Meeting BIOT division 2018 (Poster)
- "Serum Antibody Target Discovery Using Pattern Tiling"
  10th Annual Amgen-Clorox Graduate Student Symposium 2017 (Oral)
- "A Brief History of Your Immune System"
  UCSB Grad Slam 2017, Semi-finalist (Oral)
- "High-throughput epitope mapping of common antigens"
  7th ICBE—International Conference on Biomolecular Engineering 2017 (Poster)
- "High-throughput epitope mapping of common antigens"

9th Annual Clorox-Amgen Graduate Student Symposium 2016 (Poster)

- "Diagnostic Biomarker Discovery for Age-Related Macular Degeneration"
  UCSB ChE 1st Year Graduate Symposium 2014 (Oral)

TEACHING EXPERIENCE

- *TA for ChE 180A and 180B: Chemical Engineering Laboratory*, UCSB,
  Spring 2016, Spring 2017, Winter 2018
  Supervised experiments and graded reports.
- *TA for ChE 132C: Statistical Methods in Chemical Engineering*, UCSB, Winter 2015
  Graded homework, graded exams, wrote answer keys, and held office hours.
- Mentored a sophomore chemical engineering undergraduate student, UCSB,
  2015-2016
- *Scientist for ScienceLine*, UCSB, Summer 2014 – Present
  Answered scientific questions from K – 12 students in local area.
  Received ScienceLine Award in the field of Physics/Chemistry/Engineering for 2014-2015, MRL ScienceLine Award for 2015-2016, Special Recognition for Commitment to ScienceLine Award for 2016-2017
- *TA for ENGRD 2190: Mass and Energy Balances*, Cornell University, Fall 2012
  Graded homework, led recitations, and held office hours.

SKILLS AND TECHNIQUES

- Bacterial cell culture
- Bacterial peptide display
- Magnetic activated cell sorting (MACS)
- Fluorescence activated cell sorting (FACS)
- Plasmid DNA isolation
- Gel electrophoresis
- PCR
- Bacterial cloning
- Microarrays
- ELISAs
- MATLAB
- Python
- Java
- C++
- Wolfram Mathematica

ABSTRACT

Development of bioinformatic tools to identify and characterize linear protein epitopes

by

Michael Louis Paull

The adaptive immune system produces antibodies to specifically target antigens. Identifying and characterizing epitopes on target antigens enables numerous medical applications such as vaccine design, diagnostic discovery, and therapeutic antibody development. We have developed computational tools that characterize epitopes using large sets of antibody-binding peptides. To identify epitopes in target proteins, we used an approach termed K-mer Tiling of Protein Epitopes (K-TOPE). In this approach, we divided protein sequences into short overlapping subsequences of length k (k-mers). Then, we defined and scored epitopes using each k-mer's enrichment in the sets of antibody-binding peptides. Using K-TOPE, we accurately identified epitopes for monoclonal and polyclonal antibodies. Next, using 250 specimens, we identified commonly targeted epitopes in nearly 3,000 viral proteins as well as two bacterial proteomes. Importantly, these epitopes agreed with previously reported results. To map antibody binding in epitopes, we developed Multiplexed Epitope Substitution Analysis (MESA). In this approach, target epitopes were divided into short overlapping k-mers. Then, each k-mer was exhaustively substituted with all amino acids. The effects of these substitutions were used to identify amino acid preferences at important binding positions in the epitopes. By applying this method to monoclonal antibodies and multiple sets of specimens, we identified binding motifs which agreed with an alternative computational approach. Finally, K-TOPE and MESA were used

to characterize epitopes and antigens in age-related macular degeneration (AMD), herpes simplex virus (HSV), and Chagas disease. We identified 42 AMD-specific epitopes, 30 HSV2-specific epitopes, and 222 Chagas-specific epitopes. Several epitopes were in validated antigens while many were novel. MESA demonstrated that generally only 4-5 positions in these epitopes were important for binding. Future application of these approaches could enhance our understanding of the role that antibodies play in disease progression.

Table of Contents

1 Introduction 1
1.1 Motivation 1
1.2 Dissertation organization 2
1.3 Mapping serum antibody repertoires using peptide libraries 3
1.3.1 Background 3
1.3.2 Protein epitope characteristics 5
1.3.3 Classification of antibody repertoire mapping approaches 6
1.3.4 Non-random microarray methods 9
1.3.5 Non-random surface display methods 10
1.3.6 Random microarray methods 11
1.3.7 Random surface display methods 12
1.3.8 Motif identification algorithms 13
1.3.9 Summary 14
1.4 Diseases with antibody-related pathology 15
1.4.1 Age-related macular degeneration (AMD) 15
1.4.2 Herpes simplex virus (HSV) 17
1.4.3 Chagas disease 18
1.5 Techniques 19
1.5.1 Magnetic selection 19
1.5.2 Flow cytometry validation 22
1.5.3 DNA sequencing 22
2 A general approach for identifying protein epitopes targeted by antibody repertoires using whole proteomes 23
2.1 Introduction 24
2.2 Results 26
2.3 Discussion 37
2.4 Materials and methods 40
2.4.1 Strains and reagents 40
2.4.2 Bacterial peptide display and sequencing 40
2.4.3 Protein databases 41
2.4.4 Selection of literature epitopes 41

2.4.5 Sequence processing 41
2.4.6 K-TOPE algorithm 42
2.4.7 Data visualization 44
2.5 Supplemental analysis of K-TOPE 44
2.5.1 Formal statement of the epitope identification problem 44
2.5.2 Evaluating the contributions of overlapping k-mers 46
2.5.3 Assessing the validity of combining adjacent k-mers to determine epitopes 47
2.5.4 Justification of conducting analysis with 5-mers 49
3 Mapping antibody binding using multiplexed epitope substitution analysis 52
3.1 Introduction 53
3.2 Results 54
3.2.1 MESA maps binding in epitopes using random peptide libraries 54
3.2.2 Determining binding motifs for monoclonal antibodies with known epitopes. 56
3.2.3 Using MESA to identify binding motifs with a single serum specimen 58
3.2.4 Identifying binding motifs using multiple serum specimens 59
3.3 Discussion 63
3.4 Materials and methods 66
3.4.1 Bacterial display and sequencing 66
3.4.2 Monoclonal antibody spike-in 66
3.4.3 Sequence processing 66
3.4.4 MESA algorithm 67
3.4.5 Identifying antibody binding motifs with MEME 70
3.5 Supplemental analysis of MESA 71
3.5.1 Effects of MESA parameter selection on binding motifs 71
3.5.2 Determining binding motifs for K-TOPE epitopes 73
4 Identification of disease-specific epitope and antigens 77
4.1 Introduction 78
4.2 Results 80
4.2.1 Age-related macular degeneration 80
4.2.2 Herpes simplex virus 85
4.2.3 Chagas disease 90
4.3 Discussion 97

4.4 Materials and methods 100
4.4.1 Strains and reagents 100
4.4.2 Screening and sequencing bacterial peptide display libraries 101
4.4.3 Protein database searches 101
4.4.4 Epitope identification 102
4.4.5 Epitope logo generation 103
4.4.6 Data visualization 103
5 Conclusions 104
5.1 Summary 104
5.1.1 Identification of linear protein epitopes 104
5.1.2 Characterization of epitope binding motifs 105
5.1.3 Identifying and characterizing disease-specific epitopes and antigens 107
5.2 Future directions 109
5.3 Overall conclusions 111
6 References 112

Table of Figures

Figure 1.1: An overview of the classification of antibody/antigen binding interactions. 5
Figure 1.2: A general flow diagram for mapping antibody repertoires. 6
Figure 1.3: Magnetic selection and flow cytometry validation. 21
Figure 2.1: K-TOPE determines epitopes by tiling proteins into k-mers. 27
Figure 2.2: K-TOPE found epitopes for antibodies with known specificity spiked into serum. 28
Figure 2.3: A comparison of histograms generated by K-TOPE when antibodies were added to serum or buffer. 29
Figure 2.4: K-TOPE identified four epitopes in the Rhinovirus A genome polyprotein. 30
Figure 2.5: Epitopes identified through proteome searches were validated using literature-reported epitopes. 36
Figure 2.6: Comparison of RRPFF-containing epitopes in EBNA1 and Protein UL84. 47
Figure 3.1: An overview of Multiplexed Epitope Substitution Analysis (MESA). 55
Figure 3.2: MESA determined binding motifs for mAbs. 57
Figure 3.3: MESA determined binding motifs for antibodies in an individual serum specimen. 59
Figure 3.4: MESA identified binding motifs that were common in multiple specimens. 60
Figure 3.5: Binding motifs of common epitopes were identified using MESA. 62
Figure 3.6: An EBNA1 binding motif was discovered using MESA. 63
Figure 3.7: Epitope logos generated for monoclonal antibodies with varying score thresholds. 71
Figure 3.8: Monoclonal antibody epitope logos generated with varying minimum enrichment threshold percentiles. 72
Figure 3.9: Epitope logos were generated for multiple specimens with varying minimum enrichment threshold percentiles. 73
Figure 3.10: Binding motifs identified for 3 antibodies of known specificity. 74
Figure 3.11: Binding motifs identified for four Rhinovirus A epitopes. 75
Figure 3.12: Binding motifs identified for three viral epitopes. 76
Figure 3.13: Binding motifs identified for three bacterial epitopes. 76
Figure 4.1: Heat map showing epitopes scores for disease and control specimens. 83
Figure 4.2: Binding motifs for 3 epitopes that had plausible autoantigens. 84

Figure 4.3: K-TOPE identified epitopes for glycoprotein G1 using HSV1 specimens and for glycoprotein G2 using HSV2 specimens. 86

Figure 4.4: Binding motifs were determined for 3 HSV1-specific epitopes and 3 HSV2-specific epitopes. 90

Figure 4.5: Binding motif determined for the TSSA epitope ENKPATGEA. 91

Figure 4.6: Binding motifs determined for three Chagas-specific epitopes. 94

Table of Tables

Table 1.1: Information about recent antibody repertoire analysis studies. 8

Table 2.1: The expected and actual membership of different epitope groups. 31

Table 2.2: The average age for each epitope group. 32

Table 2.3: A collection of 29 viral epitopes to which >30% of 250 specimens bound. 33

Table 2.4: Epitopes in the proteomes of the genera Staphylococcus and Streptococcus which were bound by >30% of 250 specimens. 35

Table 2.5: Top 25 k-mers which were highly correlated with preceding and following k-mers. 48

Table 2.6: Library coverage for k-mers of varying length. 49

Table 2.7: The expected number of sequences for different k-mer lengths. 51

Table 4.1: AMD-specific epitopes were identified. 82

Table 4.2: Epitopes that were only bound at the final AMD timepoints. 84

Table 4.3: Alignment of an HSV2-specific glycoprotein G2 epitope with previously reported epitopes. 86

Table 4.4: HSV2-specific epitopes were identified. 87

Table 4.5: HSV2-specific epitopes in plausible antigens. 88

Table 4.6: HSV1-specific epitopes were identified. 88

Table 4.7: Eight HSV2-specific epitopes were also in the HSV1 proteome. 89

Table 4.8: Alignment of a TSSA epitope with previously reported epitopes. 91

Table 4.9: The 25 Chagas-specific epitopes with the highest prevalence. 92

Table 4.10: Comparison between SerImmune motifs and K-TOPE epitopes. 93

Table 4.11: K-TOPE epitopes matched a diagnostic panel of peptides. 94

Table 4.12: Candidate autoantigens for Chagas disease. 96

Table 4.13: Chagas-specific epitopes that do not cross-react with L. major. 97

1 Introduction

1.1 Motivation

Antibodies are produced by the adaptive immune system to target pathogens. Since antibodies specifically bind to a single target, they have found numerous uses in biological sciences and engineering. Additionally, since antibodies are a key arm of the immune response, there is interest in understanding their role in the progression of disease. There are numerous cases in which serum antibodies either indicate disease, exacerbate illness, or protect against infection. For instance, the presence of antibodies towards viruses, bacteria, or parasites can be indicative of an infection. In the case of autoimmune diseases, an autoantibody targeting self-tissues could be related to the exacerbation of a disease. Finally, vaccines generate protective antibodies and are therefore routinely administered to prevent infections. In all these cases, the ability to determine the presence of antibodies in a subject's serum is vital.

For many applications, it is important to know the target of an antibody (the antigen) and where it binds on its target (the epitope). Therefore, rather than focusing on the antibodies themselves, it is often sufficient to focus on antigens. Some methods analyze the structure of antigens to determine epitopes [1]. However, without experimental antibody binding data, the antigens targeted by a specific individual cannot be determined. Experimental methods to determine epitopes generally select antibody-binding peptides from peptide libraries [2]. Large random peptide libraries can effectively mimic epitopes and therefore bind numerous antibodies. By starting with a random library, epitopes for any protein antigen can be identified regardless of whether the antigens are from bacteria,

viruses, parasites, fungi, or animals. Screening these libraries with serum results in large sets of antibody-binding peptides which encode epitope information. Parsing these rich datasets to identify epitopes and their corresponding antigens can be challenging. An algorithm which could consistently meet this challenge would enable the determination of an "immune history" from an individual's serum. This immunological record would contain epitopes and antigens corresponding to active and past antibody responses. With the immune histories of many individuals, epitopes and antigens could be identified to enhance our ability to understand, diagnose, and treat various diseases.

1.2 Dissertation organization

This dissertation describes the development and application of algorithms that identify and characterize linear antibody epitopes. While there are research groups that routinely generate antibody-binding peptide datasets, there is a lack of computational schemes that can comprehensively identify linear epitopes *and* connect them to antigens. For our approach, we identified antibody-binding peptides using a random peptide library and processed these peptide sequences into short k-mers. These k-mers were then manipulated to characterize linear epitopes for a variety of protein antigens. To validate these methods, epitopes were generated for common pathogens as well as three diseases with antibody-related pathology.

Chapter 1 of this dissertation discusses recent studies that mapped antibody repertoires using peptide libraries, describes three model disease systems, and explains relevant experimental techniques. Chapter 2 describes the development and initial validation of the K-mer Tiling of Protein Epitopes (K-TOPE) algorithm. This algorithm uses k-mers from antibody-binding peptides and candidate antigen sequences to determine linear

epitopes. In Chapter 3, Multiplexed Epitope Substitution Analysis (MESA) is introduced, which uses k-mers to map antibody binding in epitopes. This method reveals the positions in an epitope that are important to binding and the amino acid preferences at those positions. In Chapter 4, these two algorithms are applied to age-related macular degeneration, herpes simplex virus, and Chagas disease. These analyses revealed several novel disease-specific antigens and epitopes. Finally, we summarize these studies and present future opportunities for applying K-TOPE and MESA in Chapter 5.

1.3 Mapping serum antibody repertoires using peptide libraries

Antibodies in blood provide a rich source of immunological information. Antibody repertoire analysis seeks to decode this information to empower the development of vaccines, diagnostics, and therapeutics. To this end, various approaches have been developed to determine epitopes using peptide libraries. Approaches have used random or proteome-derived peptide libraries in a microarray or surface display format. For methods using random libraries, motif discovery software has been developed to identify common binding signatures. The analysis of thousands of samples and dozens of diseases has shown that there are often disease-specific epitopes, even though individual antibody repertoires are unique. The recent developments in antibody repertoire analysis hold the potential to enable comprehensive immune evaluations.

1.3.1 Background

Antibodies bind specifically to their targets and are therefore relevant to numerous scientific, medical, and industrial applications. The immune system continuously makes antibodies even after an infection has been resolved. Therefore, the antibodies in serum

constitute an immunological record. The ability to access and interpret this record could impact many areas of biotechnology and healthcare. In particular, the identification of disease-associated antigens has enabled the development of numerous diagnostic tests for infectious [3], autoimmune [4], and allergic conditions [5]. Furthermore, epitope information can be used to inform the development of more efficacious vaccines [6]. With the growing number of antibody-based therapeutics, there is a need to characterize antibody binding to measure specificity and avoid undesired cross-reactivity [7]. Additionally, knowledge of antibody binding sites can enable the design of more effective affinity reagents for diverse applications [8]. Finally, antibody repertoire analysis methods will augment efforts to characterize the "healthy" antibody repertoire which could be useful for detecting the onset of disease [9]. With the recent development of high-density peptide microarrays, high-throughput sequencing, and increased computational power, there is increasing interest in antibody repertoire analysis.

Although the expression "antibody repertoire" is frequently used, it is informative to divide its usage into methods focusing on paratopes or epitopes. The terms paratope and epitope refer to the binding regions of the antibody and antigen, respectively (Figure 1.1). Paratope-focused methods analyze antibody CDR regions through B-cell DNA sequencing and LC-MS/MS [10,11]. These methods have proven useful for monitoring the evolution of an immune response, determining which antibody clonotypes are most abundant, and investigating class switching [12]. Alternatively, methods that identify protein epitopes have the distinct advantages of requiring minimal serum, rather than B cells, and allowing for the identification and analysis of antigens. This review focuses on approaches that use peptide libraries to determine protein epitopes for the antibody repertoire.

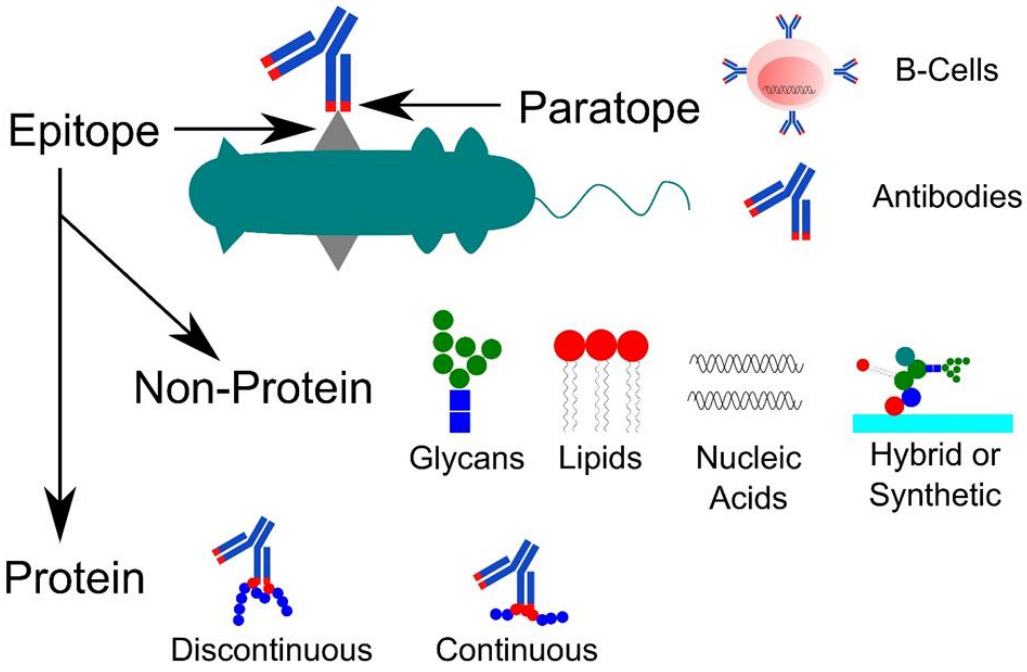


Figure 1.1: An overview of the classification of antibody/antigen binding interactions. The paratope and epitope refer to the binding regions of the antibody and antigen, respectively. Paratope-focused methods focus on serum antibodies or antibody-producing B-cells. Epitopes can be comprised of non-protein molecules such as glycans, lipids, nucleic acids, combination structures, or synthetic molecules. Continuous protein epitopes are comprised of a single sequence whereas discontinuous protein epitopes are comprised of amino acids distant in sequence, but close in the folded protein.

1.3.2 Protein epitope characteristics

Protein epitopes are typically considered to be continuous ("linear") or discontinuous ("conformational") (Figure 1.1). Continuous epitopes are comprised of a single sequence whereas discontinuous epitopes are comprised of amino acids distant in sequence, but close in the folded protein. For the cases in which the epitopes of interest are discontinuous, several methods have been developed [1,13,14]. It has been suggested that because >90% of epitopes are discontinuous, searching for continuous epitopes may be fruitless [15]. However, from an analysis of PDB antibody/antigen structures, it was determined that epitopes are generally composed of around 15 residues and that 85% of epitopes contain at

least one 5 amino acid contiguous stretch [16,17]. Strictly speaking, nearly all epitopes are discontinuous, however, the frequent occurrence of linear segments suggests that there is utility in identifying linear epitopes [18]. And importantly, linear protein epitope discovery remains bioinformatically tractable.

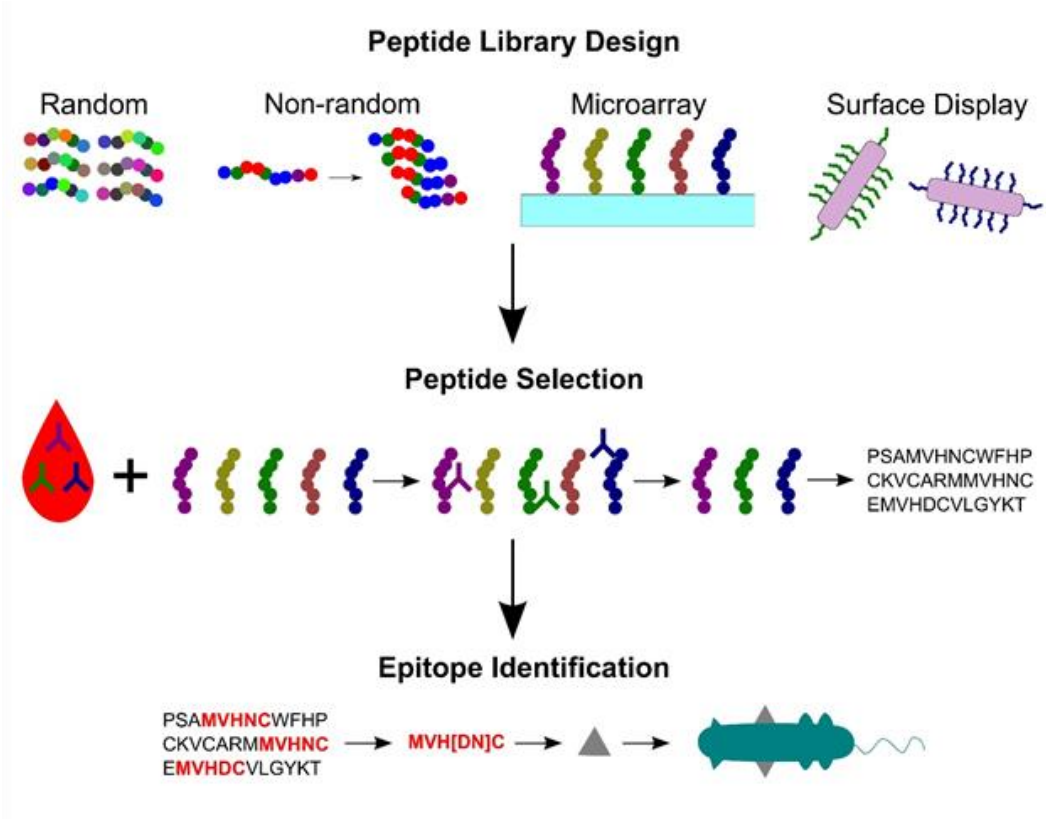


Figure 1.2: A general flow diagram for mapping antibody repertoires. First, peptide libraries are designed using random or non-random libraries displayed in a microarray or surface display format. Next, the peptide library is combined with serum and antibody-binding peptides are selected. Finally, bioinformatics is used to identify epitopes and associate epitopes with antigens using sets of antibody-binding peptides.

1.3.3 Classification of antibody repertoire mapping approaches

Approaches for mapping the antibody repertoire use random or non-random peptide libraries in a microarray or surface display format (Figure 1.2) [19]. For microarrays, peptides are typically printed or synthesized on glass slides, whereas for surface display,

peptides are most often displayed on bacteriophages, bacteria, or yeast. Relevant information for each approach referenced in this review can be found in Table 1.1.

Non-random methods often use libraries constructed by tiling target protein sequences into overlapping peptides. While the concept of determining epitopes by tiling antigens seems straightforward, there are important limitations. There is reason to believe that, paradoxically, random peptides may be able to capture binding specificities that the tiled antigens themselves cannot [20]. One potential explanation for this observation could be that an antibody binds to a peptide from a random library with higher affinity than the corresponding fragment of the antigen. Further, proteome-derived peptide libraries are typically several orders of magnitude smaller than random libraries (e.g. 10^5 [21] vs. 10^{11} [22]).

The use of random libraries allows for a less biased experimental approach to epitope determination. However, with random libraries, the burden shifts to computational motif discovery since it is necessary to resolve randomness into a coherent signal. Additionally, associating an epitope with an antigen becomes a challenge.

Table 1.1: Information about recent antibody repertoire analysis studies. The sample size, sample classification (disease/condition studied), library size, and year of publication are listed for the studies discussed. For studies using multiple libraries, the library size refers to the largest library. The sample size corresponds to the experiments using the largest library and includes both experimental and control groups. For studies with humans and animals, human sample sizes were used. In some cases, parameters were ambiguous and had to be inferred from the manuscripts.

Category	Sample size	Sample Classifications	Library Size	Year	Reference
Non-Random Microarrays	30	Multiple sclerosis	4E+03	2016	[23]
	19	Chagas disease	2E+05	2015	[24]
	10	Multiple sclerosis, narcolepsy	2E+06	2017	[25]
	22	Polyclonal antibodies	2E+06	2012	[26]
	5	Polyclonal antibodies	2E+06	2014	[8]
Non-Random Surface Display	3	Paraneoplastic syndromes	4E+05	2011	[27]
	569	Multiple viral infections	9E+04	2015	[28]
	61	Dengue virus	4E+03	2017	[29]
	298	Multiple sclerosis, type 1 diabetes, rheumatoid arthritis	4E+05	2013	[21]
Random Microarrays	11	Alzheimer's disease	1E+04	2012	[30]
	61	Pancreatic diseases	1E+04	2012	[31]
	125	Multiple viral infections, multiple cancers	3E+05	2014	[32]
	42	Myalgic encephalomyelitis	1E+05	2016	[33]
	1516	Multiple viral infections, multiple cancers	1E+04	2014	[34]
	106	Multiple viral and bacterial infections, malaria	3E+05	2015	[35]
	73	Healthy controls	3E+05	2016	[9]
Random Surface Display	1	Human immunodeficiency virus	2E+09	2012	[36]
	5	High-grade epithelial ovarian cancer	1E+11	2016	[22]
	14	Human immunodeficiency virus	1E+09	2013	[37]
	10	Vaccinations	1E+09	2013	[38]
	2	Healthy controls	1E+09	2017	[39]
	12	Peanut allergies	1E+09	2015	[40]
	88	Celiac disease	1E+10	2012	[41]
	38	Pre-eclampsia	8E+09	2016	[42]
	29	Celiac disease	8E+09	2016	[43]

Microarrays are potentially more reproducible, less laborious, and more quantitative than surface display systems. However, microarrays usually display 3-5 orders of magnitude fewer peptides than surface display libraries, which can attain library diversities up to 10^{11}. A consequence is that microarrays may not contain enough information for certain applications. Also, surface display peptide libraries can be propagated by growth which reduces cost.

1.3.4 Non-random microarray methods

For non-random microarrays, antigens from a pathogen of interest or the human proteome are tiled into overlapping peptides. Multiple Sclerosis (MS) autoantibodies have been examined in depth using a microarray with presumed MS autoantigens and Epstein-Barr Virus (EBV) antigens [23]. This analysis discovered peptides that were bound by the serum antibodies of MS subjects, but not by matched controls. A peptide containing "RRPFF" from EBV Nuclear Antigen 1 (EBNA1) was stated to be disease-specific, however a study with a larger sample size showed that this epitope was prevalent in the general population [21]. Antigens of the parasitic protozoan *Trypanosoma cruzi* have been tiled to analyze serum samples from subjects infected with Chagas disease [24]. In this case, the use of high-density microarrays identified multiple new disease-specific peptides and antigens. However, only a fraction of *Trypanosoma cruzi* antigens could be examined because of the large size of this parasite's proteome. These examples demonstrate that microarray size can restrict the number of serum samples and antigens that can be analyzed.

Another method used a human proteome microarray with six amino acid lateral shifts followed by a targeted microarray with only single amino acid shifts [25]. This method identified two potential novel autoantigens for narcolepsy and multiple sclerosis. An approach for determining the fine specificity of epitopes used a non-random microarray,

followed by an exhaustive mutagenesis scheme on selected epitopes [26]. The scheme was later refined and made available through the online server ArrayPitope [44].

1.3.5 Non-random surface display methods

An exemplar surface displayed non-random library is T7-Pep, in which peptides representing the human proteome are incorporated into a phage library [27]. Human proteome libraries are useful for probing autoantibodies using phage immunoprecipitation sequencing (PhIP-Seq). The initial application of this method identified candidate autoantigens in subjects with paraneoplastic syndromes. T7-Pep was also used for a large-scale PhIP-Seq screen of nearly 300 antibody repertoires from subjects with type 1 diabetes, multiple sclerosis, rheumatoid arthritis, and healthy controls [21]. Most antibody-binding peptides were unique to individuals, suggesting that each antibody repertoire is unique. Even with large sample sizes, disease-specific peptides with high sensitivity were not found. Their absence may be due to disease heterogeneity or the inherent stochasticity of the humoral response.

One of the largest antibody repertoire studies to date screened 569 antibody repertoires using a phage library with peptides tiling virtually all human-host viruses [28]. By including known infected serum specimens, researchers could identify virus-specific epitopes. This study found that there were “public epitopes” bound by the antibodies of many subjects. There were also notable differences in the epitopes bound by subjects of varying ages and geographic locations. One limitation of this method was that it used 56 residue peptides which makes it difficult to localize an epitope to its core 5-10 amino acids. Also, vaccine-related epitopes for common viruses such as measles and rubella were not observed

in this study [45]. A smaller scale study used a library derived from the Dengue virus proteome and found shared epitopes in infected subjects [29].

1.3.6 Random microarray methods

Immunosignaturing uses 10^4-10^5 random peptides displayed in a microarray format to profile the antibody repertoire [46–48]. Immunosignaturing has been used to profile the antibody repertoires of transgenic mice with an Alzheimer's disease (AD) phenotype to identify distinct signatures at different time points in disease progression [30]. Interestingly, these results suggested that humans with AD had detectable immunological similarity despite having distinct personal antibody repertoires. In other studies, immunosignatures differentiated between similar pancreatic diseases [31], between multiple cancers and infectious diseases [32], and between myalgic encephalomyelitis disease subjects and controls [33]. Microarrays were also used to successfully classify more than 1500 serum specimens into 15 disease groups [34]. Since many diseases and specimens were simultaneously and successfully classified, this lends support to the idea that a single peptide microarray could potentially identify many distinct diseases.

Subsequences have been used instead of sequences in BLAST searches to identify potential associations between epitopes and pathogen antigens [35]. Researchers determined that in a BLAST search of pathogen proteins, the true antigen can be resolved if a pair of pentamers exactly match the antigen or a pair of heptamers have 80% identity. However, this strategy makes the restrictive assumption that an antigen of interest has multiple significant epitopes. To address the changes in the antibody repertoire in "healthy" humans, immunosignaturing was used on longitudinal serum samples and determined that a person's

immunosignature remains fairly constant over time [9]. Despite this consistency, signatures corresponding to vaccinations administered during the study were observed.

1.3.7 Random surface display methods

A method that exemplifies random surface display is Deep Panning, which uses a large random phage library and Next-Generation Sequencing (NGS) [36]. In this case, the clustering algorithm MEME was used to identify motifs within the set of selected peptides [49]. Phage display and NGS have also been used to identify tumor-associated antigen peptides for high-grade epithelial ovarian cancer [22]. To identify epitopes associated with HIV, rhesus macaques were vaccinated against HIV and vaccination-specific epitopes were found using biopanning [37].

Studies have shown that motifs typically require seven fixed amino acids to identify specific antigens within the entire non-redundant database using BLAST [50,51]. Although, if database searches were restricted to specific proteomes or if multiple motifs could be matched to a protein, it was possible to reach statistical significance [38]. A combination approach was employed which identifies antibody-binding peptides using a random phage-displayed library, assays these peptides in a microarray format, and uses an additional microarray for substitution analysis [39]. Peptides were clustered using the MEME algorithm and the resultant motifs were analyzed using BLAST to identify candidate antigens within common pathogens.

Phage display and NGS were used to investigate IgE epitopes from subjects with peanut allergies [40]. Selected peptides were aligned to a known peanut allergen and pairwise clustering was used to determine motifs. Interestingly, this study demonstrated that subject-

specific motifs could be observed in early selection rounds. This observation suggests that motifs could be identified with fewer selection rounds, which could help avoid selection bias.

Peptide libraries displayed on bacteria have also found utility in antibody repertoire analysis [52,53]. Bacterial display libraries of random peptides have been used to identify a panel of diagnostic peptides for celiac disease [41]. Similarly, bacterial display was used to identify disease-associated epitopes in subjects with pre-eclampsia [42]. Molecular mimicry was suspected since a prominent consensus motif was linked to EBNA1, which in turn exhibited similarity to a human protein, GPR50. This demonstrates that antibody repertoire analysis could also focus on the environmental causes of autoimmune diseases brought on by molecular mimicry.

1.3.8 Motif identification algorithms

A significant limitation of the commonly employed motif discovery algorithm, MEME, is that it scales approximately quadratically with the number of input sequences. It is therefore necessary to reduce a list of antibody-binding peptides to less than about 5000 sequences. This is often accomplished through additional selection rounds which can detrimentally remove relevant peptides. In an effort to address the limited throughput of MEME, the MUSI algorithm was developed which can quickly identify multiple motifs in tens of thousands of sequences [54]. This method represents a time complexity improvement over MEME, though it is not clear how well it would scale to datasets of greater than 10^5 sequences, which are routinely generated using NGS. Gibbs clustering has also been used to discover motifs in large datasets [55]. This algorithm has greater accuracy than MUSI, but may be slower.

The anchor based sequence clustering algorithm (ASC) focuses on a preliminary clustering step, followed by using existing algorithms to identify motifs within the clusters [56]. This algorithm outperformed previous motif discovery algorithms such as MEME, MUSI, and Gibbs clustering. Another algorithm that discovers motifs in large datasets is IMUNE, in which antibody-binding peptides are reduced to a set of enriched patterns which are then clustered together to determine motifs [43]. Using IMUNE, disease-specific motifs for celiac disease were identified. Since IMUNE reduces large sets of peptides to patterns, the algorithm's runtime is approximately linear with respect to the initial number of peptides.

1.3.9 Summary

Antibody repertoire analysis methods hold the potential to substantially increase our understanding of adaptive immune responses. Various approaches have been developed that will likely fill different niches. Microarray-based methods offer potential benefits in terms of ease of use and assay time. However, approaches using surface display may be better for epitope and antigen discovery since they offer significantly larger libraries. Tiled libraries could be useful for testing specific hypotheses about suspected antigens. Although, for the exploration of a large variety of antibody-antigen interactions, random libraries offer more flexibility.

To analyze large datasets, it is generally useful to reduce the peptide sequences to subsequences or motifs. This transformation is performed by Richer et. al. [35], MUSI [54], Gibbs Clustering [55], anchor based sequence clustering [56], and IMUNE [43]. Mapping epitopes to their corresponding antigens when using random libraries remains a significant challenge. Motifs discovered in peptide datasets are often too short to definitively connect to antigens using a protein database search.

Antibody repertoire analysis will continue to fuel the discovery of disease-specific epitopes and the development of diagnostic assays. In addition, these approaches could inform the development of therapeutics for cancer, infections, and autoimmune diseases. Antibody repertoire analysis may also aid efforts to develop vaccines and quantitatively measure their efficacy. We are approaching a point at which we will have the capability to consistently access this hidden cache of immunological information.

1.4 Diseases with antibody-related pathology

Three diseases with prominent antibody responses are age-related macular degeneration (AMD), herpes simplex virus (HSV), and Chagas disease. Analyzing these antibody responses could aid the discovery of epitopes that indicate disease. Also, identifying the targets of antibody responses could reveal disease etiologies, which would aid therapeutic and vaccine development.

1.4.1 Age-related macular degeneration (AMD)

Age-related macular degeneration (AMD) is the leading cause of blindness in the developed world [57]. Ageing is the key risk factor for AMD, since the immune system becomes increasingly dysregulated with age [58]. This dysregulation leads to an increased susceptibility to autoimmune diseases and the generation of self-reactive autoantibodies. AMD causes a progressive loss of central vision due to atrophy of the retinal cells, which is known as the “dry” or atrophic form, or abnormal blood vessel growth, which is referred to as the “wet” or neovascular form [59]. Around 90% of AMD cases are the less severe dry form, but these cases can lead to the accelerated vision loss of the wet form [60]. There are currently 10 million people with vision impairment as a result of AMD, with 2 million people

that are blind [61]. Globally, the prevalence of AMD for ages 45-49 years is 4.2%, but jumps to 27.1% for ages 80-84 years [62].

In AMD, death of the retinal pigment epithelial (RPE) cells and photoreceptors occurs in the macula, which is a small area in the back of the eye that is responsible for the majority of useful vision [63]. The RPE cells are responsible for nourishing photoreceptors as well as transporting away their metabolic waste [64]. Disruption of these duties is related to the accumulation of extracellular deposits called drusen between the RPE cells and an adjacent basement membrane complex called Bruch's membrane [65]. The correlation between the presence of drusen and the early stages of AMD is sufficiently strong that the presence of drusen is used as one of the main diagnostic criteria for AMD [66]. There is no cure for AMD and effective treatment exists only for the wet form and consists of photodynamic therapy and anti-vascular endothelial growth factor drugs [67].

A mounting body of evidence suggests that the immune system plays an integral role in the progression of AMD [63,68–70]. This role has been demonstrated by the presence of complement proteins in drusen, the presence of anti-retinal antibodies, and the increased risk of AMD in patients with an elevated number of memory T-cells [71]. A key finding for AMD was that a single nucleotide polymorphism in the immune regulatory protein Complement Factor H (CFH) led to a 2.5- to 6.0-fold increase in AMD risk [72]. This finding was used to develop a monoclonal antibody drug that targets Complement Factor D, lampalizumab, for the treatment of dry AMD [73]. Notably, autoantibodies against human proteins have been found in the sera of AMD patients [74–81]. Efforts to elucidate the autoantibody targets that contribute to the pathology of AMD have focused on LC-MS/MS

[78,81] and protein microarrays [80]. There would be clear utility in identifying reliable serological biomarkers for the diagnosis of AMD [59,82].

1.4.2 Herpes simplex virus (HSV)

Herpes simplex virus 1 (HSV1) and herpes simplex virus 2 (HSV2) cause human infections in the orofacial region ("cold sores") and the genital region (genital ulcers), respectively [83]. HSV is characterized by its ability to reactivate in the present of humoral immunity after a period of latency [84]. This latency along with an elaborate system for blocking host immune responses increases the pathogenicity of HSV. In addition to mild recurrent labial or genital lesions, HSV can cause keratoconjunctivitis, visceral HSV infections in immunocompromised hosts, HSV encephalitis, and lethal neonatal infections [84]. In 2012, the global prevalence of HSV1 was 3.7 billion people ages 0-49 [85] and the global prevalence of HSV2 was 417 million people ages 15-49 [86]. HSV2 has synergy with HIV, increasing the risk of HIV-acquisition by three-fold, increasing transmissibility by five-fold, and accelerating HIV disease progression [86]. Thus, current research efforts are generally focused on developing HSV2-specific diagnostics.

HSV1 and HSV2 contain the same genes [87] and the protein-coding regions of the HSV1 and HSV2 genomes share 83% sequence homology [88]. The only striking instance of a protein that varies significantly between the two viruses is glycoprotein G [89]. Since glycoprotein G is the optimal candidate for identifying HSV2-specific diagnostics, it has been a frequent target in epitope mapping [90–92]. As a result of these analyses, glycoprotein G and its epitopes have been used in diagnostics for HSV2 [93,94]. Several envelope glycoproteins have been characterized as potential diagnostics since they are accessible to antibodies on the exterior of the virus [95–100]. Recently, both HSV1- and HSV2-specific

epitopes have been identified by using a microarray to analyze the glycoproteins [101]. Thus, identifying epitopes for the glycoproteins or even the whole proteomes of HSV1 and HSV2 could aid in the development of effective diagnostics.

1.4.3 Chagas disease

Chagas disease is a chronic infection by the parasitic protozoan *Trypanosoma cruzi (T. cruzi)* that can lead to cardiomyopathy and digestive megasyndromes [102]. The protozoan is primarily transmitted to humans by large, blood-sucking insects called triatomines ("kissing bugs") [103]. In addition to parasite-specific responses, heart-specific autoimmune responses have been identified for Chagas disease [104]. This disease affects 8 million people in Latin America [102], with an estimated incidence of 4% per year in endemic regions [105]. Chagas disease is the most important parasitic disease in the western hemisphere with a disease burden of 7.5 times that of malaria [105]. This burden makes Chagas disease the leading cause of cardiac lesions in young, economically productive adults in Latin America [106]. While the acute phase of Chagas disease resolves spontaneously in 90% of individuals, 30-40% of patients develop a chronic form with cardiac and digestive pathology 10-30 years after the initial infection [102]. Thus, it is important to have an effective diagnostic for Chagas disease to treat patients before serious symptoms develop [107,108].

The genome of *T. cruzi* has been fully sequenced [109], enabling the analysis of multiple *T. cruzi* antigens. Epitopes have been identified for trypomastigote small surface antigen (TSSA) [24,110,111], which plays an important role in infectivity [112]. This antigen has already found utility as a Chagas diagnostic [113]. Additional studies have identified Chagas-specific epitopes using a proteome-derived microarray [24], a random microarray

[114], and a random surface-displayed library [115]. Multiple *T. cruzi* antigens have been identified including mucin TcMUCII [114,116], trans-sialidase [117], dispersed gene family protein 1 (DGF-1) [114,118], surface antigen 2 (B13) [119], and mucin-associated surface protein (MASP) [120–122]. Importantly, since the causative agent for leishmaniasis, *Leishmania major (L. major)*, has 85% cross-reactivity with *T. cruzi* [119], diagnostics need to differentiate between Chagas disease and leishmaniasis [123]. By including epitopes from known *T. cruzi* antigens in an ELISA, researchers diagnosed Chagas disease with high sensitivity and specificity against leishmaniasis [116]. Taken together, these studies suggest that it would be impactful to identify novel Chagas-specific epitopes.

1.5 Techniques

1.5.1 Magnetic selection

An in-depth protocol for the experimental techniques used in this dissertation can be found in a recent study [124]. Briefly, a large, random 12-mer peptide library with approximately 8 X 10^9 members was displayed as part of the transmembrane protein eCPX on *E. coli*. The main goal of these experiments was to isolate cells displaying antibody-binding peptides. This was accomplished by combining human serum with the peptide library and sorting for antibody-binding library members.

One possible concern with this approach is that antibodies could bind to the *E. coli* cell surface directly, rather than to the displayed peptides. To avoid this scenario, serum was depleted of *E. coli* binding antibodies. The serum depletion was performed by first combining serum and cells displaying only the protein eCPX. All antibodies that bound cells directly were removed, leaving the depleted serum.

Once serum was depleted, the bacterial library was sorted using magnetic selection (Figure 1.3A). To magnetically label cells that display antibody-binding peptides, we used protein A/G magnetic beads. These beads bound to the constant region of the antibodies, whereas the peptides bound to the variable region of the antibodies. To prepare a "cleared library" of cells that do not bind directly to the beads, peptide-displaying cells were combined with beads. Then, a magnet was applied to remove cells that bound directly to the beads. Next, we combined the depleted serum with the cleared library, added magnetic beads, and applied a magnet to sort for cells displaying antibody-binding peptides. These sorted cells were denoted "the enriched population". Finally, the entire magnetic selection procedure was repeated, except starting with the enriched population rather than the initial library.

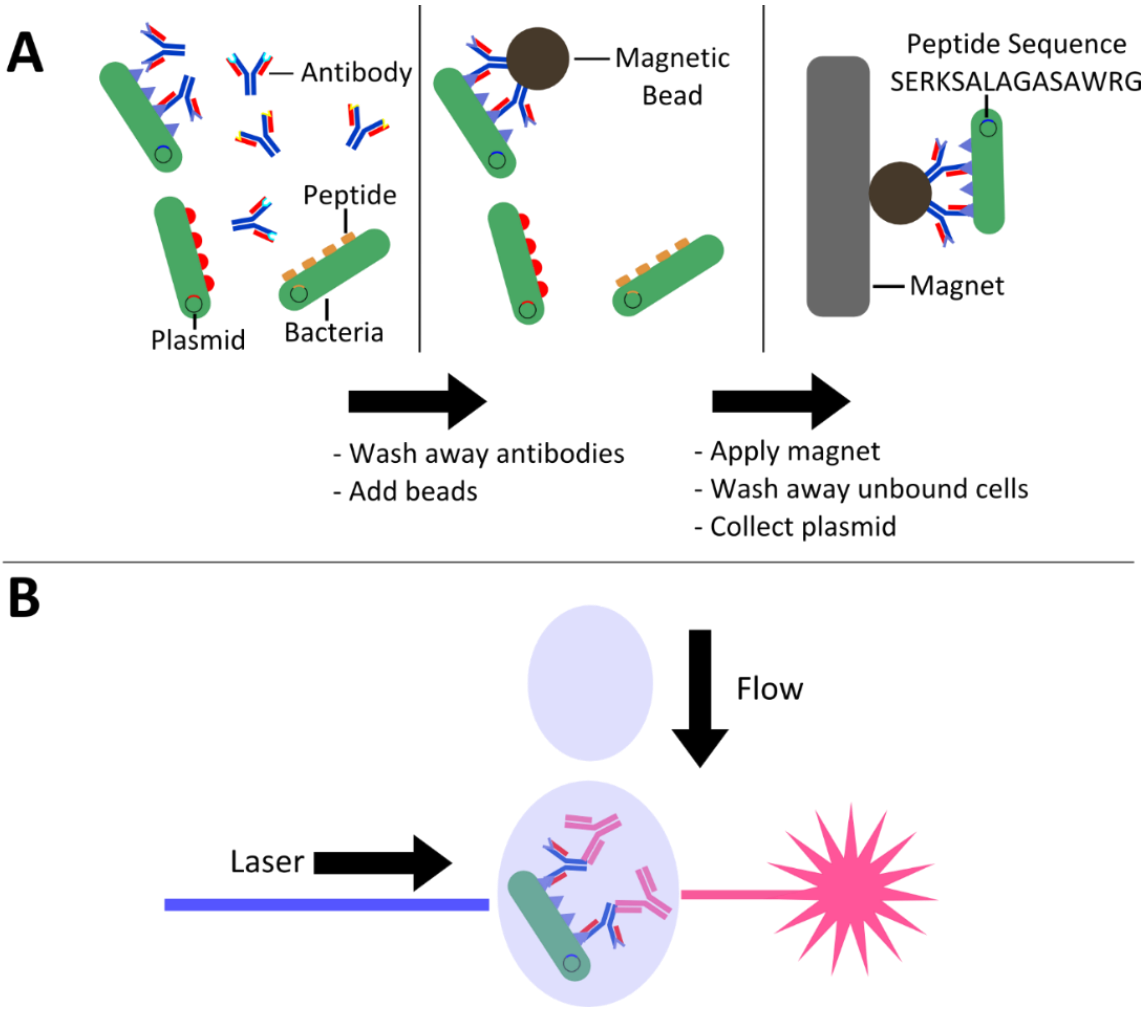


Figure 1.3: Magnetic selection and flow cytometry validation. (A) We combined peptide displaying bacteria with serum and then washed away unbound antibodies. Next, we added magnetic beads which bound to the constant region of antibodies. We applied a magnet to capture cells with bound antibodies and then washed away unbound cells. The plasmids from this enriched population of bacteria were collected and sequenced. (B) We combined the enriched population of bacteria with serum and a fluorescent secondary antibody. The cells were examined with a laser such that cells with a higher quantity of bound antibodies had higher intensity fluorescence. This measurement revealed the percentage of cells in the enriched population that bound antibody.

1.5.2 Flow cytometry validation

We utilized flow cytometry to confirm that a majority of the enriched population could bind antibodies (Figure 1.3B). In flow cytometry, a laser is directed at cells suspended in small droplets to determine the fluorescence of individual cells. To start, the enriched population was combined with serum and a fluorescent secondary antibody that bound to the constant region of IgG antibodies. Cells displaying only the protein eCPX were used to establish the background level of expression. Then, we measured the fluorescence of the enriched library above the background level of expression. Generally, if the fluorescence above background was greater than 50%, the library was prepared for sequencing. Often, the enriched population had 80-90% fluorescence above background.

1.5.3 DNA sequencing

To extract DNA, we lysed the enriched population cells and collected the plasmids. Then, we used PCR to amplify the sequences in the plasmids coding for peptides. To multiplex specimens, we assigned a separate barcode to each serum specimen. Next, we sequenced the barcoded DNA using the NextSeq 500, usually generating approximately 300-500 million reads. This translated to approximately 50 million total sequences for 30 specimens. These DNA sequences were then translated to amino acid sequences. Finally, we combined sequences that only differed due to sequencing errors.

2 A general approach for identifying protein epitopes targeted by antibody repertoires using whole proteomes

Antibodies are essential to functional immunity, yet the epitopes targeted by antibody repertoires remain largely uncharacterized. To aid in characterization, we developed a generalizable strategy to identify antibody-binding epitopes within individual proteins and entire proteomes. Specifically, we selected antibody-binding peptides for 250 distinct sera out of a random library (10^{10} members) and identified the peptides using next-generation sequencing. To identify antibody-binding epitopes and the antigens from which these epitopes were derived, we tiled the sequences of candidate antigens into short overlapping subsequences of length k (k-mers). We used the extent to which each of these k-mers was enriched over background in the antibody-binding peptide dataset to identify antibody-binding epitopes. As a positive control, we used this approach, termed K-mer Tiling of Protein Epitopes (K-TOPE), to identify epitopes targeted by monoclonal and polyclonal antibodies of well-characterized specificity, accurately recovering their known epitopes. To characterize a commonly targeted antigen from *Rhinovirus A*, K-TOPE identified three epitopes recognized by antibodies present in 83% of sera (n = 250). An analysis of 2,908 proteins from 400 viral taxa that infect humans revealed seven enterovirus epitopes and five Epstein-Barr virus epitopes recognized by >30% of specimens. Analysis of *Staphylococcus* and *Streptococcus* proteomes similarly revealed six epitopes recognized by >40% of specimens. These common viral and bacterial epitopes exhibited excellent agreement with previously mapped epitopes. The K-TOPE approach thus provides a powerful new tool to elucidate the organisms, antigens, and epitopes targeted by human antibody repertoires.

2.1 Introduction

Immunological memory allows for rapid antibody responses towards diverse antigens long after initial exposure. For example, the adaptive immune response to many vaccinations is often sustained throughout an individual's lifetime [125]. This immunological information is archived within the genes encoding B-cell and T-cell receptors along with the corresponding receptor structures, but has proven difficult to characterize in a comprehensive manner. The ability to more fully interrogate immunological memory could reveal exposures to pathogens, commensal organisms, and allergens. Such information has proven useful for correlating antibody responses with disease outcomes to design more effective vaccines [6]. A detailed record of immune exposures can also facilitate the identification of biomarkers to diagnose infectious [3], autoimmune [4], and allergic conditions [41]. Furthermore, the capability to broadly characterize antibody repertoires at the epitope level could be used to identify conserved pathogen epitopes [42] and tumor specific antigen epitopes [126] to aid in therapeutic discovery.

Immunological memory has been investigated extensively through sequencing the variable regions of B- and T-cell receptor encoding genes amplified from circulating cells [10]. These methods have proven useful for identifying receptor-encoding genes that associate with vaccination [127]. Nevertheless, such genetic information has not generally provided insight into the specific environmental antigens and epitopes targeted, unless they are known *a priori*. Furthermore, these methods require large specimen volumes (>10 mL) to obtain a sufficient quantity of cells [127]. Thus, there remains a need for methods that identify the diverse antigen targets of adaptive immunity.

Several methods have been developed to profile the protein epitopes of the secreted antibody repertoire [2]. One common approach to epitope mapping is to generate short overlapping peptides by tiling candidate antigens. These peptides are then assayed for serum antibody reactivity in peptide microarray [23] or bacteriophage display library [28] formats. However, because these methods are biased towards specific organisms, they do not enable comprehensive or hypothesis-free immune evaluation. One strategy to overcome the limitations of tiling experiments is to use fully random peptide libraries [32,38,41]. Here, experiments are less biased and methods can analyze epitopes corresponding to a variety of organisms and antigens. A disadvantage of microarrays is that they are typically several orders of magnitude less diverse than peptide display libraries (e.g. 10^5 [32] versus 10^{10} [41]), limiting the effectiveness with which current methods can achieve epitope discovery for low titer antibodies. In random library experiments, epitopes are typically discovered using *de novo* motif discovery by unsupervised clustering [49]. The most widely used algorithm for this purpose, MEME, scales approximately quadratically with the number of input sequences, making it less useful for analyzing large datasets resulting from next generation sequencing (NGS). While full-length antibody-binding peptides can be analyzed, the majority of the binding energy is typically derived from just 5-6 amino acids [46], thus other amino acids within the peptide will contribute noise. To rectify this problem researchers developed the IMUNE algorithm to reduce peptide datasets into statistically enriched patterns and cluster these patterns to build motifs [43].

A significant challenge for epitope mapping approaches is the association of epitopes and motifs with their corresponding antigens. Typically about seven amino acids need be specified to unambiguously identify the corresponding antigen within the full database of

protein sequences [51]. Because linear stretches in epitopes are typically less than seven amino acids in length [17], protein database searches of individual epitopes (such as through BLAST [50]) often fail to achieve statistical significance. Using multiple epitope matches within a single candidate antigen can increase the confidence of antigen prediction [35,38]. However, this method is insufficient for antigens with a single important epitope. To address this challenge, we present a general approach for associating epitopes with antigens using large peptide datasets. The K-mer Tiling of Protein Epitopes (K-TOPE) algorithm identifies epitopes by computationally tiling candidate antigens into k-mers, which are then evaluated within large datasets of antibody-binding peptides. Here, we demonstrate the utility of this approach by identifying linear epitopes within several prevalent infectious pathogens.

2.2 Results

To enable the identification of protein epitopes bound by serum antibodies, we developed a method that uses a database of antibody-binding peptides to identify epitopes in known protein sequences (Figure 2.1). First, we selected peptides binding to an individual antibody repertoire within a specimen (serum or plasma) from a bacterial display peptide library with 10^{10} random 12-mer members. Then, we identified antibody-binding peptide sequences using NGS. To allow for the manipulation of 20^5 (3.2 million) k-mers rather than full-length peptides, we processed peptides into subsequences and evaluated the enrichments of all k-mers of length 5 [43]. Next, K-TOPE tiled candidate antigen sequences, such as from a proteome, into overlapping k-mers. K-TOPE used the enrichment values for these k-mers to construct an enrichment histogram across the length of each protein sequence. The value at each sequence position in the histogram was proportional to the enrichment of k-mers that included that position. Epitopes were extracted from the maxima in the histogram and scored

based on their area under the curve (AUC). Finally, epitopes were assigned an "epitope percentile" based on their rank in a list of scores generated from random proteins.

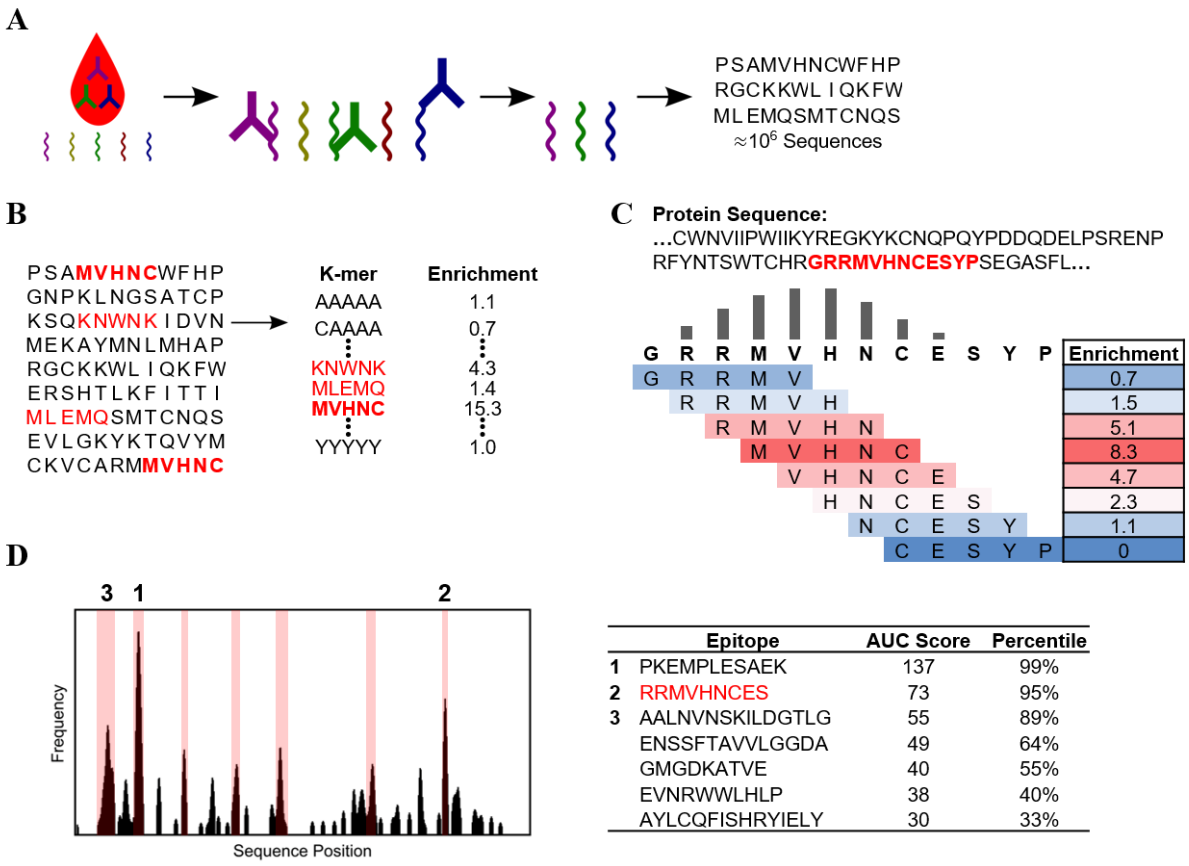


Figure 2.1: K-TOPE determines epitopes by tiling proteins into k-mers. (A) The input to the algorithm is a dataset of approximately 10^6 peptides that were bound by serum antibodies. (B) All 5-mers are evaluated for their enrichment in the list of peptides. (C) A portion of a protein sequence is tiled into 5-mers which are weighted by their enrichment. This determines a "frequency" value for each position in the sequence. (D) The frequency value for each position in a protein sequence is plotted as a histogram. Possible epitopes are highlighted in pink on the graph. Epitope sequences, area under the curve (AUC) scores, and significance percentiles are displayed.

To assess the utility of K-TOPE, we first determined epitopes for monoclonal and polyclonal antibodies that bind specific, well-defined epitopes in cMyc, V5, and amyloid beta. We spiked these antibodies into serum at a final concentration of 25 nM and then

selected and identified binding peptides. K-TOPE identified epitopes that corresponded closely to the previously reported epitopes of these antibodies (Figure 2.2). Importantly, the enrichment histograms generated by antibodies spiked into background serum or buffer were nearly identical (Figure 2.3), suggesting that the noisy serum environment minimally affected epitope identification.

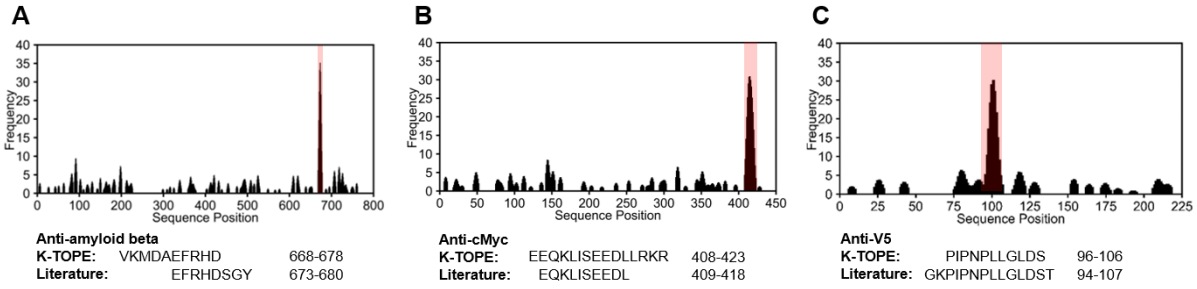


Figure 2.2: K-TOPE found epitopes for antibodies with known specificity spiked into serum. Histograms for antibodies with known specificity against amyloid beta (P05067), cMyc (P01106), and V5 (P11207) had prominent epitopes (in pink). (A) K-TOPE analysis of amyloid beta determined the epitope VKMDAEFRHD (668-678). This antibody was raised to whole protein and is known from literature to have a conformation-specific discontinuous epitope that maps to segments EFRHDSGY (673-680) and ED (692-693). (B) K-TOPE analysis of cMyc determined the epitope EEQKLISEEDLLRKR (408-422). This antibody was raised to AEEQKLISEEDLLRKRRE (407-424). (C) K-TOPE analysis of V5 determined the epitope PIPNPLLGLDS (96-106). The antibody was raised to GKPIPNPLLGLDST (94-107).

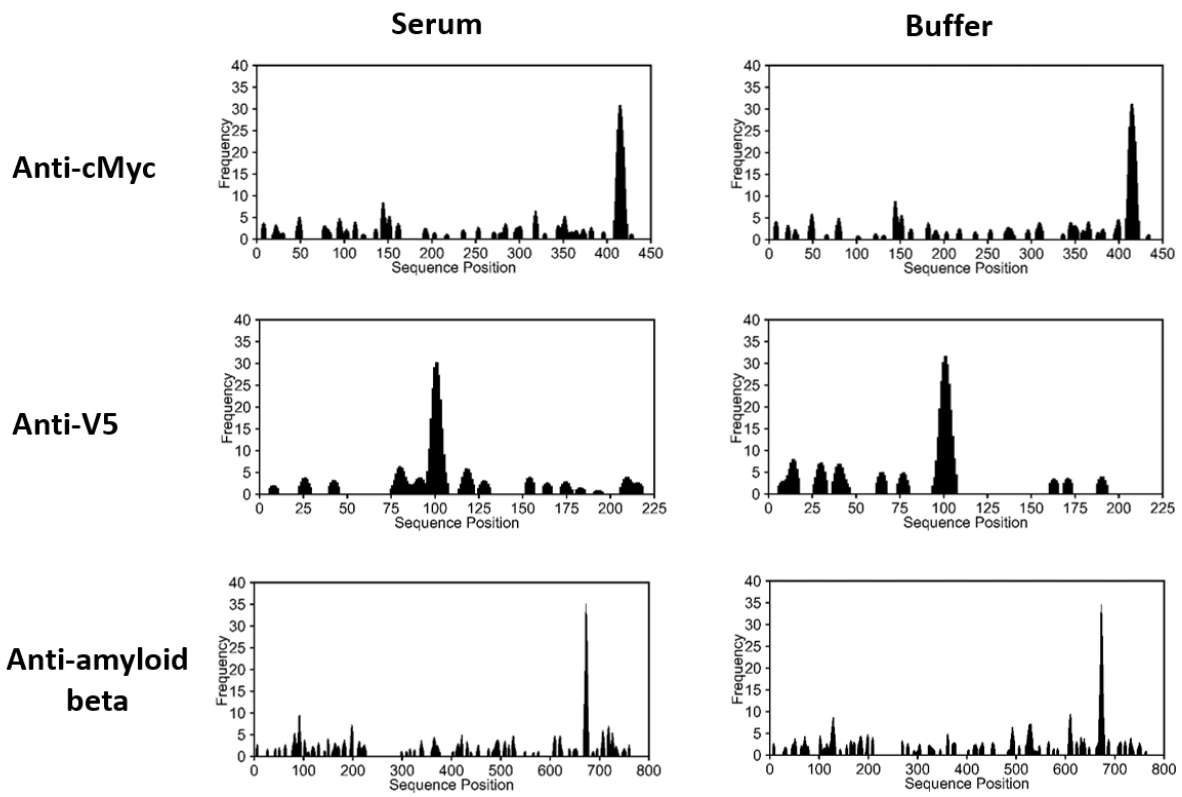
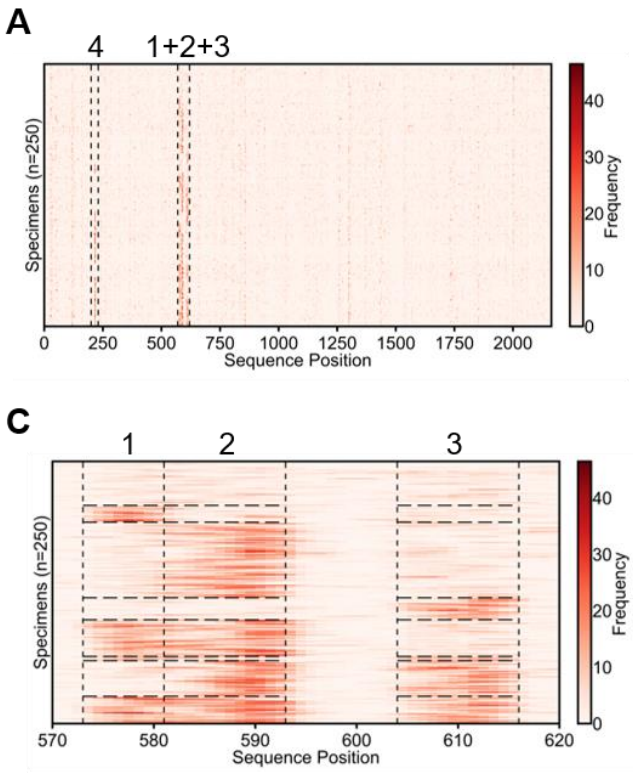


Figure 2.3: A comparison of histograms generated by K-TOPE when antibodies were added to serum or buffer. Histograms were generated for antibodies against cMyc (P01106), V5 (P11207), and amyloid beta (P05067). The most prominent peaks were present regardless of whether antibodies were added to serum or buffer. This suggests that the binding signature of a single antibody was not obscured by the many other antibody specificities present in serum.

To identify "public epitopes" conserved across many individuals, epitopes were generated for each specimen individually and then clustered. Given the ubiquity of exposure to the upper respiratory pathogen *Rhinovirus A*, we validated the approach by identifying epitopes within its genome polyprotein. Using a unique set of 250 serum specimens, we identified epitopes within *Rhinovirus A* that were targeted by 30% or more of the specimens (Figure 2.4A). Of the 250 specimens, 87% exhibited binding to at least one of these consensus epitopes (Figure 2.4B). Three of these epitopes were located within positions 570-

620 (Figure 2.4C), in the antigenic attachment region of VP1. A fourth epitope within the VP2 region of the *Rhinovirus A* genome polyprotein was targeted by 43% of the population.



Epitope #	% Population Binds
1	32%
2	67%
3	34%
4	41%
Any of the 4	87%

Figure 2.4: K-TOPE identified four epitopes in the *Rhinovirus A* genome polyprotein. (A) K-TOPE was applied to the *Rhinovirus A* genome polyprotein (P07210) for 250 specimens. Histograms for all specimens are shown as rows in a heat map. The specimens have been clustered such that specimens that bind the same epitopes are adjacent. Regions that contain epitopes are outlined by dotted lines. (B) A table of the percentage of the population that bound each epitope. For instance, Epitope 1 is the percentage of specimens that targeted "1", "1+2", "1+3", "1+4", "1+2+3", "1+2+4", "1+2+3+4". (C) The region from positions 570-620 is divided into 3 sections that correspond to distinct epitopes. These epitopes are consensus epitopes which were present in >30% of the 250 specimens. (D) Bar graph showing membership in different epitope groups. For example, a specimen that binds epitopes 2 and 3 will belong to epitope group "2+3". In this population, 87% of the specimens bound at least one of the consensus epitopes. The sequences of the epitopes were 1: QNPVENYI, 2: DSVLEVLVVPN, 3: APALDAAETGHT, and 4: NHTHPGEQG.

To assess trends in the population, each specimen was assigned into one of eight groups based on which of the three VP1 epitopes were bound (Figure 2.4D). Notably, epitope binding was not independent, since the group of specimens targeting all three epitopes was 44% larger than expected and the group targeting epitopes '1+3' was 50% smaller than expected (Table 2.1). The average age of the subset of specimens of known age (n=138) was 35 years, however, the epitope group targeting all three epitopes had an average age of 17, and the epitope group targeting none of the epitopes had an average age of 50 (Table 2.2). Thus, people who targeted fewer *Rhinovirus A* epitopes tended to be older.

Table 2.1: The expected and actual membership of different epitope groups. The expected membership of epitope groups was calculated by multiplying the proportions of the population that bound each epitope. For example, if epitope 1 was bound by 32% of the population and epitope 2 was bound by 67%, then the expected membership of epitope group '1+2' would be 21%. Note that specimens in groups *only* bound the epitopes in the groups e.g. specimens in group '1' did not bind '2' or '3'. Generally, the actual and expected membership values agreed except for the '1+2+3' group which had higher membership than expected and the '1+3' group which had lower membership than expected (in bold).

Group	Actual	Expected	Percent Difference
1	16	17	-6%
2	72	75	-4%
3	21	19	11%
1+2	35	35	0%
1+3	**4**	**8**	**-50%**
2+3	34	38	-11%
1+2+3	**26**	**18**	**44%**
None	42	37	14%

Table 2.2: The average age for each epitope group. The average age for the 138 specimens for which there was age data was 35. The 'None' group had an average age of 50 which was notably higher than the average age of 35 (in bold). Additionally, the '1+2+3' group had a lower average age of 17 (in bold). This discrepancy suggests that older people targeted fewer *Rhinovirus A* epitopes.

Group	Average Age	Group Size
1	29 ± 24	6
2	23 ± 25	35
3	37 ± 23	11
1+2	21 ± 21	14
1+3	39 ± 18	2
2+3	26 ± 24	20
1+2+3	**17 ± 12**	**11**
None	**50 ± 22**	**19**
All Specimens	**35 ± 27**	**138**

Next, we investigated the utility of using K-TOPE to identify epitopes within a set of 2,908 proteins from 400 viral taxa with human tropism. This approach yielded 29 epitopes that were bound by at least 30% of all specimens (Table 2.3). The prevalence of each epitope is noted, which is defined as the proportion of specimens that bound the epitope. Some of these epitopes have been reported previously [21,42,128,129], while a few were likely due to false discovery (e.g., Mayaro virus and Lyssavirus). Thus, a modest number of prominent linear viral epitopes were bound by >30% of the specimens analyzed. A common antigen identified from this analysis was Epstein-Barr nuclear antigen 1 (EBNA1) from Epstein-Barr virus (EBV), which is expressed in EBV-infected cells [130]. Additionally, the epitopes identified for the enterovirus genus were consistent with the epitopes identified for *Rhinovirus* A, which is a species in that genus (Figure 2.4).

Table 2.3: A collection of 29 viral epitopes to which >30% of 250 specimens bound. K-TOPE was used to analyze 2,908 proteins from viruses with human tropism. This search demonstrated that only a few prominent linear viral epitopes were bound by a large portion of the population.

Epitope	Protein	Taxon	Accession	Prevalence
DSVLNEVLVVPN	Genome polyprotein	Enterovirus	P07210	0.668
PALTAAETG	Genome polyprotein	Enterovirus	Q66575	0.588
GRRPFFHPV	Epstein-Barr nuclear antigen 1	Epstein-Barr virus (strain GD1)	Q1HVF7	0.524
AGAGGGAGA	Epstein-Barr nuclear antigen 1	Epstein-Barr virus (strain GD1)	Q1HVF7	0.516
KYTHPGEA	Genome polyprotein	Enterovirus	Q82122	0.492
VRRPFFSD	Protein UL84	Human cytomegalovirus	P16727	0.452
NPVERYVDE	Genome polyprotein	Enterovirus	Q82122	0.428
MVVPEFK	DNA-binding protein	Human mastadenovirus C	P03265	0.428
EVKLPHWTPT	Glycoprotein 42	Epstein-Barr virus (strain GD1)	P03205	0.42
KPQPEKPK	Structural polyprotein	Mayaro virus	Q8QZ72	0.416
GGAGAGGAGAGGG	Epstein-Barr nuclear antigen 1	Epstein-Barr virus (strain GD1)	P03211	0.412
ININRPLE	Large structural protein	Lyssavirus	Q9QSP0	0.412
RPSCIGCKG	Epstein-Barr nuclear antigen 1	Epstein-Barr virus (strain GD1)	P03211	0.404
GAGAGAGGG	Packaging protein UL32	Simplexvirus	P89455	0.376
LEEVIVEKTK	Genome polyprotein	Enterovirus	Q82081	0.352
KHTHPGI	Replication origin-binding protein	Human herpesvirus 3	P09299	0.352
AETGHTNKI	Genome polyprotein	Enterovirus	Q82122	0.344
YVFPHWITK	Envelope glycoprotein gp63	Primate T-lymphotropic virus 3	Q0R5Q9	0.34
KTTNTTTNT	Immediate-early protein 2	Roseolovirus	Q9QJ16	0.34
MAADKPTL	Genome polyprotein	Murray Valley encephalitis virus	P05769	0.34
SFIVPEFA	Virion membrane protein A16	Orthopoxvirus	P16710	0.332
LVLPHWYMA	Cytoplasmic envelopment protein 1	Simplexvirus	P89430	0.328
YVDDMLNDI	Large tegument protein deneddylase	Human herpesvirus 6A (strain Uganda-1102)	P52340	0.328
SSGPKHTQKV	Genome polyprotein	Enterovirus	P03303	0.324
PVPEFQA	Non-structural polyprotein	Semliki forest virus	P08411	0.316
VPVTPNIAI	Genome polyprotein	Hepatitis C virus	Q68749	0.304
LHRPALTA	Minor capsid protein L2	Human papillomavirus type 34	P36758	0.304
EHILNRPTG	RNA-directed RNA polymerase L	Crimean-Congo hemorrhagic fever orthonairovirus	Q6TQR6	0.304
GEFIGSE	Shutoff alkaline exonuclease	Human herpesvirus 8	Q2HR95	0.3

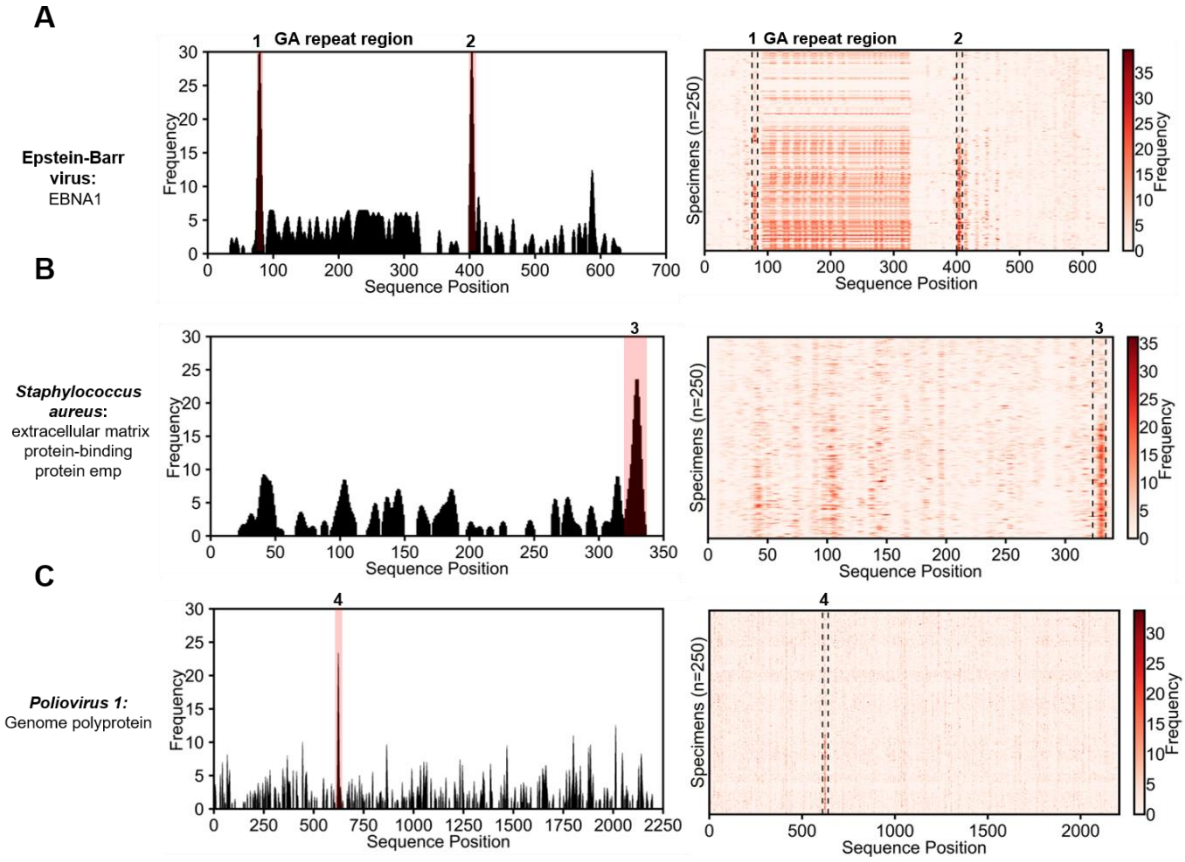
We performed a similar analysis for the proteomes of the genera *Streptococcus* and *Staphylococcus*, which are common bacterial human pathogens with 2,976 and 3,071 proteins in their respective proteomes. K-TOPE was used with each of these proteomes to determine epitopes bound by >30% of a population of 250 specimens, yielding 9 epitopes for *Streptococcus* and 13 epitopes for *Staphylococcus* (Table 2.4). The epitope LIPEFIG(R) in ATP-dependent Clp protease ATP-binding subunit ClpX was the most prevalent *Streptococcus* epitope and second most prevalent *Staphylococcus* epitope. Therefore, K-TOPE could not determine which genus generated this epitope. The most prevalent *Staphylococcus* epitope was PTHYVPEFKGS from extracellular matrix protein-binding protein emp, which is a known virulence factor [131]. For *Streptococcus*, the second most prevalent epitope was GQKMDDMLNS from the highly antigenic Streptolysin O protein [132]. This epitope falls within a 70 amino acid range in Streptolysin O that is known to bind antibodies [133]. The sequence "DKP" was present in 5/9 *Streptococcus* epitopes and the sequence "PEFXG" was present in 6/13 *Staphylococcus* epitopes (Table 2.4). Therefore, there are multiple candidate antigens that may correspond to these highly enriched sequences.

Table 2.4: Epitopes in the proteomes of the genera *Staphylococcus* and *Streptococcus* which were bound by >30% of 250 specimens. K-TOPE was used to analyze 2,976 proteins from *Streptococcus* and 3,071 proteins from *Staphylococcus*.

Epitope	Protein	Accession	Prevalence
Streptococcus			
LIPEFIGR	ATP-dependent Clp protease ATP-binding subunit ClpX	P63793	0.512
GQKMDDMLNS	Streptolysin O	Q5XE40	0.436
QIPALDKPL	FMN-dependent NADH-azoreductase	A4W2Z7	0.416
IADKPILD	UPF0154 protein SSU05_1707	A4VX34	0.392
TVADKPVA	Phenylalanine--tRNA ligase beta subunit	Q5XCX3	0.360
RTPDKPT	Agglutinin receptor	P16952	0.324
VVPNIWR	Putative 2-dehydropantoate 2-reductase	P65666	0.320
LLNRPIHD	CCA-adding enzyme	Q5M153	0.320
TLADKPEF	Autolysin	P06653	0.308
Staphylococcus			
PTHYVPEFKGS	Extracellular matrix protein-binding protein emp	Q2FIK4	0.572
LIPEFIG	ATP-dependent Clp protease ATP-binding subunit ClpX	B9DNC0	0.508
NKPEFSGAT	3-isopropylmalate dehydratase small subunit	Q4L7U3	0.436
NKNNKNNKN	Translation initiation factor IF-2	Q4L5X1	0.372
KLGNIVPEYK	Extracellular matrix protein-binding protein emp	P0C6P1	0.360
KLCRICFRE	30S ribosomal protein S14 type Z	Q5HM12	0.352
DFLNRPVD	Proline--tRNA ligase	Q4L5W5	0.348
EKNNNNNNNNS	Alkaline shock protein 23	Q4L860	0.320
GVVPNISR	UvrABC system protein A	Q5HHQ9	0.312
LIPEFNQV	Homoserine kinase	Q8CSQ2	0.308
SPEFLGSQ	Undecaprenyl-diphosphatase	B9DK59	0.308
VGINRPTY	Putative glycosyltransferase TagX	O05154	0.308
VIPEFNND	Peptide chain release factor 2	Q4L4H9	0.300

The most prevalent epitopes identified through proteome searches were validated by comparison to previously reported epitopes. We chose to analyze the viral proteins EBNA1 from EBV and the *Poliovirus 1* genome polyprotein (representing Enterovirus), which were present five and seven times, respectively, in Table 2.3. Bacterial proteins chosen for validation were Streptolysin O, corresponding to the second most prevalent *Streptococcus* epitope (Table 2.4), and Extracellular matrix protein-binding protein emp, corresponding to most prevalent *Staphylococcus* epitope (Table 2.4**)**. In all cases, K-TOPE found prominent peaks in the histograms that corresponded to reported epitopes (Figure 2.5) [21,39,42,129]. Additionally, K-TOPE identified an immunogenic region of GA-repeats from positions 100-350 in the analysis of EBNA1 [23]. We used a nonparametric statistical test to assign

significance to the overlap between K-TOPE epitopes and known epitopes. Using this method, all epitopes evaluated using K-TOPE had P-values below 0.05 (Figure 2.5D).



Peak #	Literature	K-TOPE	Prevalence	P-Value	References
1	KRPSCIGCK	RPSCIGCKG	0.40	0.0056*	[42]
2	RRPFF	PGRRPFFHP	0.52	0.0031	[21]
3	VPEFKGSLP	PTHYVPEFKGS	0.57	0.019	[39]
4	PALTAVETGATNPL	EIPALTAVETG	0.39	0.003	[129]

Figure 2.5: Epitopes identified through proteome searches were validated using literature-reported epitopes. In (A), (B), and (C), a histogram is shown for a single specimen with significant peaks (in pink). To the right of the histogram is a heat map for 250 specimens. For (A), there is a region of antigenic GA-repeats from positions 100-350. The table in (D) provides the statistical significance of agreement between literature epitopes and K-TOPE epitopes for the labeled peaks in (A), (B), and (C). The UniProt accessions used for this analysis were P03211 for EBNA1, Q8NXI8 for extracellular matrix protein-binding protein emp, and P03300 for *Poliovirus 1* Genome Polyprotein. Statistical tests where epitopes with >50% GA content were removed are denoted by an asterisk "*". All identified epitopes had p-values below 0.05.

2.3 Discussion

Here, we present a generalizable methodology for identifying epitopes within candidate immunogenic proteins. By tiling proteins into k-mers and evaluating those k-mers in a database of antibody-binding peptides, we determined epitopes for individuals and a population. One of the main features of this approach is that it combines k-mers to determine composite epitopes that may not explicitly exist in the peptide dataset. Another important element is using an antigen sequence to identify epitopes, thereby surmounting the 7 amino acid requirement for successful antigen identification [51].

The K-TOPE approach to epitope mapping differs from reported methods in several important ways. First, using a library that contains peptides spanning all viral proteomes cannot easily be extended to much larger bacterial or parasitic proteomes [28]. A disadvantage of microarrays is that they have far lower 5-mer coverage (~27% [35]), than surface display (~100%) which could limit the application of k-mer approaches. Other algorithms have been developed that identify binding motifs in peptide datasets, but they lack the integrated capability to connect motifs to protein antigens [54,56]. Also, the direct method of aligning peptides to sequences becomes computationally infeasible with a large number of peptides and candidate antigens [134].

The heterogeneity of experimental approaches complicates the validation of putative epitopes and their associated antigens. The Immune Epitope Database (IEDB) has an all-inclusive representation of information [135], which may not reflect important distinctions in experimental platforms, specimens, and data analysis techniques. For instance, there are likely numerous false positive epitopes for highly studied organisms and few identified epitopes for less studied antigens. Also, there is a lack of quantitative data reported for

epitopes [136], such as the proportion of a given population that binds an epitope. To address this lack of information, we used K-TOPE to analyze specimens for responses to common pathogens in a general population. This allows newly identified "public epitopes" to be benchmarked by nearly any set of serum specimens. We determined public epitopes in *Rhinovirus A* and showed that people who targeted fewer *Rhinovirus A* epitopes tended to be older, perhaps due to immunosenescence [137], reduced pathogen exposure, or a lower incidence of rhinovirus infections [138] . With a diverse group of specimens, it was possible to confirm that the RRPFF epitope in EBV's protein EBNA1 is a very commonly targeted epitope [21], rather than a multiple sclerosis-specific epitope [23]. Since the specimens used in this study were not assayed for responses to pathogens, acute and chronic infections could not be readily distinguished from prior infections. It will be valuable to validate this method using antibody-binding peptides that reflect acute infection. Additionally, longitudinal studies could aid the identification of epitopes which appear upon vaccination or acute infection [9].

K-TOPE provides a new tool for identifying diagnostic biomarkers, vaccine components, and candidate therapeutic targets. This approach could be used in the iterative process of designing a vaccine, since it would be useful to know which epitopes are elicited in a population by vaccination. Vaccine formulation could be altered to maximize the percentage of the population that targets epitopes associated with a positive disease outcome [6]. K-TOPE could also enable the development of diagnostics that assign disease based on the presence of epitopes. Since this method only involves a single experimental screen, in principle multiple diseases could be simultaneously diagnosed [34]. By searching for consensus epitopes in a disease group that are absent in a control group, K-TOPE could

discover disease-specific epitopes. For an autoimmune disease, the entire human proteome could be analyzed to determine autoantigen epitopes [21]. Similarly, using clinical histories of viral infection, K-TOPE could analyze the proteomes of suspected pathogens to link epitopes to infections [28]. With specimens that have HLA information, it could be possible to detect a correlation between HLA type and bound epitopes [139]. This could have implications for how we determine genetic predisposition to immunological disease.

There are important limitations to the conditions in which this approach could be successful. First, this approach is currently limited to the identification of linear epitopes. However, since 85% of epitopes have at least one linear stretch of five amino acids [17], conformational epitopes with linear segments may be represented in the datasets. This report focuses on epitopes from common pathogens which are high-titer, but it could be difficult to detect rare antibody epitopes. Methods that selectively deplete out high-titer antibodies could prove effective for probing rare antibodies [140]. Another limitation is that protein sequences tend to have a large degree of conservation and redundancy [141], as demonstrated by the Enterovirus epitope PALTAVETGATNPL [129]. As another example, the *Human herpesvirus 6A* epitope YVDDMLNDI (Table 2.3) contains "DDMLN", which is shared by the *Streptococcus* epitope GQKMDDMLNS (Table 2.4). Generally, if an epitope sequence is present identically in multiple antigens, all candidate antigens should be considered equally plausible without further biological or epidemiological information.

In summary, the present approach enables the discovery of epitopes within the proteomes of any organism whose sequence is deposited into the protein database. The challenge of associating epitopes with antigens can be surmounted by transforming sets of antibody-binding peptides to k-mers and tiling proteins of interest. Advancements upon this

paradigm may enable comprehensive immunological evaluations from serum and other biological tissues.

2.4 Materials and methods

2.4.1 Strains and reagents

E. coli strain MC1061 was used with surface display vector pB33eCPX for all library screening experiments. Protein A/G magnetic beads were from Thermo Scientific Pierce. Antibodies with known specificity included C3956 rabbit anti-c-Myc polyclonal antibody (Sigma), anti-beta amyloid 1-42 antibody [mOC31] - conformation-specific (ab201059) (Abcam), and rabbit V8137 Anti-V5 polyclonal antibody (Sigma). Antibodies were spiked into healthy donor serum at a concentration of 25 nM. All sera (n=251) were obtained as deidentified specimens from biobanks according to institutional guidelines, (Biosafety authorization numbers #201417, #201713), and handled according to CDC-recommended BSL2 guidelines.

2.4.2 Bacterial peptide display and sequencing

The bacterial peptide display screening protocol was carried out as previously described [43,124]. Briefly, an *E. coli* library displaying approximately 8 billion different 12-mer peptides was combined with 1:100 diluted serum. We used magnetic selection with Protein A/G beads to isolate bacterial cells with bound antibodies. Then, we confirmed that this isolated fraction of bacteria bound antibodies using flow cytometry. Amplicons were prepared from the isolated fraction for sequencing using the Illumina NextSeq.

2.4.3 Protein databases

Protein sequences were obtained from UniProt or by using the Biopython module [142]. Accessions for proteins are noted in figures and figure captions. For the epitope validation, accessions were chosen that reference the most highly annotated version of the proteins identified in Table 2.3 and Table 2.4. The list of random proteins used for statistical analysis was obtained through a UniProt search of “reviewed:yes”. The viral proteome search used a Uniref search of “uniprot:(host:"homo sapiens" reviewed:yes fragment:no) AND identity:0.9” and yielded 2,908 proteins. The *Staphylococcus* proteome search used a Uniref search of “uniprot:(taxonomy:"Staphylococcus [1279]" fragment:no reviewed:yes) AND identity:0.9” and yielded 3,071 proteins. The *Streptococcus* proteome search used a Uniref search of “uniprot:(taxonomy:"Streptococcus [1301]" fragment:no reviewed:yes) AND identity:0.9” and yielded 2,976 proteins.

2.4.4 Selection of literature epitopes

For EBNA1, RRPFF was chosen because it was noted that RRPFF antibodies occur with equal frequency in the serum of MS and healthy individuals [21]. KRPSCIGCK was noted as an EBNA1 epitope that was preferentially targeted by pre-eclamptic women, but was also targeted by healthy controls [42]. The motif XPEFXGSXX was discovered and inferred to correspond to VPEFKGSLP in *Staphylococcus aureus* using protein database searches [39]. For *Poliovirus 1,* the epitope PALTAVETGATNPL was found to be a cross-reactive epitope in many enteroviruses [129].

2.4.5 Sequence processing

The algorithms for generating nonredundant sequence lists from FASTQ files, outputting enrichment values for subsequences, and exhaustively calculating k-mer statistics

were adapted from IMUNE [43]. We added the capability to start with lists of peptides rather than NGS data. The enrichment of a k-mer is defined as the ratio of the number of observations of the k-mer to the "expected" number of observations. The "expected" value is calculated as the product of the total number of sequences, the number of frames the k-mer could fit in the sequences, and the probability of the k-mer appearing based on amino acid usage. If a k-mer's enrichment is above the "enrichment minimum" (2.0 for this study), it is used in K-TOPE. K-mers need to be calculated only once per specimen. All interaction with IMUNE-derived code is through a Python module which sets up a folder hierarchy and acts as a wrapper for IMUNE-derived code. These programs are memory and hard-drive intensive and it is recommended to have at least 16 GB of free RAM and 100 GB of hard-drive space. Analysis was carried out on a Dell Optiplex 9020 with an Intel® Core™ i7-4790 CPU @ 3.60 GHz, 64-bit operating system, and 32.0 GB of RAM. Processing FASTQ files into subsequences from 12 specimens, each containing approximately 1.5 million unique sequences, required 2.3 hours and calculating k-mer enrichment required 7.7 minutes. The duration of these calculations scales approximately linearly with the number of specimens and sequences.

2.4.6 K-TOPE algorithm

The K-TOPE algorithm is written in Python 3.6 and will be available online. First, there is a RAM-intensive step of loading k-mer enrichment data into memory as a dictionary. The enrichment dictionary for 250 specimens required approximately 4 GB of RAM. Then, a protein of interest is chosen for analysis and its sequence is loaded. This protein is tiled into k-mers of a set length. For this study, 5-mers were used. Each position in the protein sequence is assigned a frequency counter that starts at 0. The frequency counter of each

sequence position contained in an enriched k-mer is incremented by the logarithm base 2 of the k-mer's enrichment. The frequency counters are compiled into a histogram which is smoothed using a moving window. For this analysis, the window had width 7 and used linear weighting with 1 in the center and 0.1 at the edges. Minima and maxima are identified in the smoothed histogram. All intervals between 2 minima that contain a maximum are used to define epitopes. Epitopes were limited to a minimum length of 6 and a maximum length of 15. Epitopes are scored using the area under the curve of the un-smoothed histogram. To assign statistical significance to each epitope, the epitope's score is ranked in a list of scores for epitopes of the same length generated through an analysis of 10,000 random proteins. This rank is reported as a percentile in the distribution of random protein epitope scores. For this study, a percentile cutoff of 95% was used. For 12 specimens, analysis of 10,000 random proteins required 10.0 minutes.

After determining epitopes for individual specimens, K-TOPE can determine consensus epitopes for a population. Each epitope is characterized by a "centroid" which is the weighted central position of the epitope, indexed as a position in the protein sequence. Centroids for all epitopes that meet the percentile cutoff are compiled. They are then clustered using k-means to associate close centroids with the KMeans function from scitkit-learn [143]. A representative epitope is made for each cluster and kept if it meets a minimum prevalence in the population. Closely overlapping epitopes are removed and the final list is sorted by prevalence. Consensus epitopes can be determined for each protein in a proteome, generating a list of epitopes prevalent in a population. Determination of consensus epitopes for the *Rhinovirus A* genome polyprotein (P07210) for 250 specimens required 24.4 seconds.

The proteome searches for viruses with human tropism, *Staphylococcus*, and *Streptococcus* for 250 specimens required 3.1, 2.3, and 1.9 hours, respectively.

We calculated expected membership of epitope groups by multiplying the proportions of the population that bound each epitope. For example, if epitope 1 was bound by 32% of the population and epitope 2 was bound by 67%, then the expected membership of epitope group '1+2' would be 21%. We ranked the overlaps between K-TOPE derived epitopes and literature epitopes in a list of 10,000 randomly generated epitope overlaps to determine a p-value. To remove redundant epitopes found in the proteome searches, we used the PAM30 similarity matrix to align two epitopes and compare each position to calculate a similarity score. Epitopes that had similarity scores >10, were in the same protein, and were from different organisms were considered redundant. We removed the less prevalent of the two redundant epitopes.

2.4.7 Data visualization

Figure 2.1 was created using Inkscape. Histograms and heat maps were generated using the Matplotlib python module [144]. Bar graphs were generated using GraphPad Prism 7 Software.

2.5 Supplemental analysis of K-TOPE

2.5.1 Formal statement of the epitope identification problem

The process of identifying epitopes with K-TOPE requires the following inputs:

$\boldsymbol{RR}$: $\{R_1, R_2, \ldots, R_n\}$ is a set of sets of antibody-binding peptides

$n \sim 1 - 250$ different specimens

$\boldsymbol{R_n}$: $\{r_1, r_2, \ldots, r_l\}$ is a set of antibody-binding peptides

$l \sim 10^6$ different peptides

$\boldsymbol{r_l}$: $\{c_1, c_2, \ldots, c_m\}$ is a string of amino acids

$m = 12$ amino acids

$\boldsymbol{P}$: $\{P_1, P_2, \ldots P_q\}$ is a set of protein sequences

$q \sim 1 - 5{,}000$ different proteins

$\boldsymbol{P_q}$: $\{c_1, c_2, \ldots, c_v\}$ is a string of amino acids

$v \sim 100 - 3000$ amino acids

Using these inputs, the following outputs are generated:

$\boldsymbol{E}$: $\{E_{nq1}, E_{nq2}, \ldots, E_{nqi}\}$ epitopes for specimen n and protein q

$i \sim 0 - 10$ epitopes

$\boldsymbol{E_{nqi}}$: $\{c_s, c_{s+1}, \ldots, c_t\}$ amino acids which are a subsequence of protein q and

meet the constraints $1 \leq s \leq t \leq r$ and $6 \leq |t - s| \leq 15$

$\boldsymbol{S}$: $\{s_{nq1}, s_{nq2}, \ldots, s_{nqi}\}$ scores for specimen n, protein q, and epitope i

$s \geq 0$

$\boldsymbol{RS}$: $\{rs_1, rs_2, \ldots, rs_v\}$ scores for epitopes generated from v random proteins

$v = 10{,}000$

$\boldsymbol{PS}$: $\{ps_{nq1}, ps_{nq2}, \ldots, ps_{nqi}\}$ percentiles for specimen n, protein q, and epitope i

Each epitope E_{nqi} has its score s_{nqi} ranked in $\boldsymbol{RS}$ to determine ps_{nqi}

$0 \leq ps \leq 100$

$\boldsymbol{CE}$: $\{ce_{q1}, ce_{q2}, \ldots, ce_{qu}\}$ consensus epitopes for protein q

$\boldsymbol{CES}$: $\{ces_{q1}, ces_{q2}, \ldots, ces_{qu}\}$ scores for epitopes in $\boldsymbol{CE}$

$\boldsymbol{CEP}$: $\{cep_{q1}, cep_{q2}, \ldots, cep_{qu}\}$ prevalence values for epitopes in $\boldsymbol{CE}$

Thus, the goal of epitope identification is to develop a mapping from the peptide set $\boldsymbol{RR}$ to the protein set $\boldsymbol{P}$ to identify epitope sets $\boldsymbol{E}$ and $\boldsymbol{CE}$**.** The epitope set $\boldsymbol{E}$ has an associated score set $\boldsymbol{S}$ and percentile set $\boldsymbol{PS}$. The consensus epitope set $\boldsymbol{CES}$ has an associated score set $\boldsymbol{CES}$ and prevalence set $\boldsymbol{CSP}$.

2.5.2 Evaluating the contributions of overlapping k-mers

Many epitopes are dominated by a single highly enriched "core" k-mer, such as in the case of "GRRPFFHPV" from EBNA1 (Figure 2.5A) where the average enrichment of "RRPFF" for 250 specimens was 299. The next most enriched k-mer in this epitope was PFFHP with an average enrichment of 87. Therefore, it is reasonable to question how much additional information is contributed by the k-mers surrounding the core k-mer. This question can be addressed by examining two epitopes that share a highly enriched core k-mer. In this case, the surrounding k-mers will determine which epitope has a higher score or prevalence. For example, the epitopes GRRPFFHPV in EBNA1 and VRRPFFSD in Protein UL84 of human cytomegalovirus (CMV) (Table 2.3) both share the core k-mer RRPFF. However, the EBNA1 epitope had a prevalence of 0.524 and the Protein UL84 epitope had a prevalence of 0.452. Therefore, the inclusion of the k-mers surrounding RRPFF led to a

higher prevalence for the EBNA1 epitope than for the Protein UL84 epitope. The average enrichment for each k-mer that composes the EBNA1 and Protein UL84 epitopes is displayed in Figure 2.6. For the EBNA1 epitope, the surrounding k-mers were more highly enriched than for the Protein UL84 epitope. We can infer that the antigen with the higher prevalence epitope is the more plausible origin of the RRPFF-containing epitope. By this argument, the more plausible source of the RRPFF-containing epitope is EBNA1. Thus, the k-mers surrounding the core k-mer contributed significant information and can be used to determine which antigens are more plausible**.**

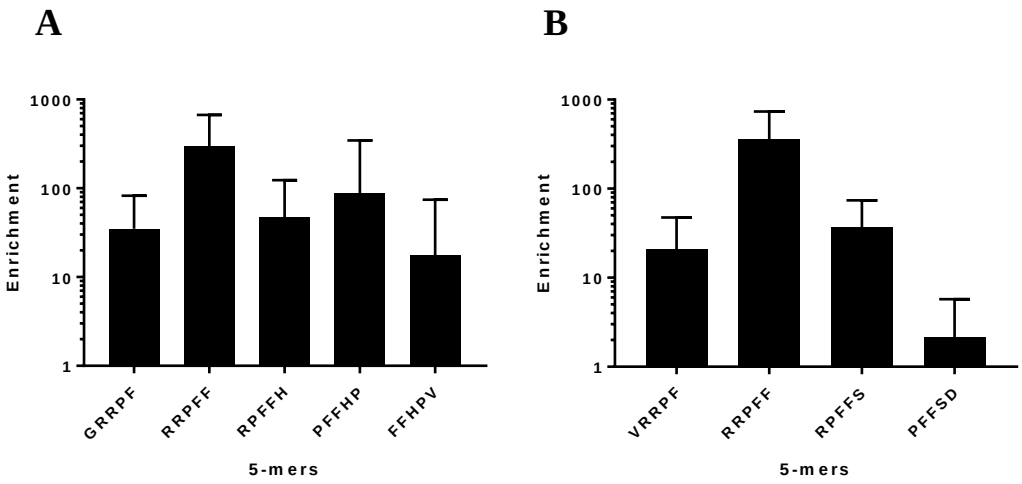


Figure 2.6: Comparison of RRPFF-containing epitopes in EBNA1 and Protein UL84. (A) The enrichment of k-mers comprising the EBNA1 epitope. (B) The enrichment of k-mers comprising the Protein UL84 epitope. The k-mers surrounding RRPFF in EBNA1 are more highly enriched than the k-mers surrounding RRPFF in Protein UL84.

2.5.3 Assessing the validity of combining adjacent k-mers to determine epitopes

To further explore the validity of combining information from overlapping k-mers, we calculated the correlation between the enrichments of adjacent k-mers. We assumed that if adjacent k-mers were similarly enriched for a set of specimens, then they were part of the same epitope. This analysis used the *Rhinovirus A* genome polyprotein, for which prevalent

epitopes were identified earlier (Figure 2.4). We tiled the *Rhinovirus A* genome polyprotein into 5-mers and calculated the enrichment of each 5-mer in 250 specimens. Then, enrichment values were normalized to 0 or 1 depending on whether the values were above the 95th percentile, ranked in the list of all 5-mer enrichments. Thus, each 5-mer was associated with a set of 250 binary scores (one for each specimen). Then, we calculated the correlation between each 5-mer's set of binary scores with all other 5-mer score sets. To determine regions of correlated 5-mers, we used a moving window approach where we averaged each 5-mer's correlation with the correlations of the three preceding and three following 5-mers. This analysis revealed 5-mers with enrichments that were highly correlated with adjacent k-mers (Table 2.5).

Table 2.5: Top 25 k-mers which were highly correlated with preceding and following k-mers. There were highly correlated regions from approximately 577-588 and 214-216.

K-mer position	Correlation Average
588	0.53
587	0.53
586	0.52
585	0.51
46	0.5
214	0.5
579	0.5
215	0.48
216	0.48
584	0.47
577	0.46
580	0.46
47	0.46
583	0.45
581	0.44

This analysis showed that there were highly correlated regions from approximately positions 577-588 and 214-216. As shown earlier (Figure 2.4), *Rhinovirus A* epitope

QNPVENYI was at positions 573-580, epitope DSVLNEVLVVPN was at positions 581-592, and epitope NHTHPGEQG was at positions 213-222. Thus, the regions of highly correlated k-mers corresponded closely with the epitopes determined using K-TOPE. The reason for this agreement is likely because these k-mers are bound by the same antibodies. Therefore, combining consecutive enriched k-mers to determine longer epitopes appears valid.

2.5.4 Justification of conducting analysis with 5-mers

We chose to use 5-mers for K-TOPE analysis because they were the optimal k-mer length based on library coverage and enrichment dynamic range. Here, library coverage of a k-mer is defined as the percentage of all 20^k k-mers that are observed at least once in the 12-mer peptide library before selection. The library coverage percentages of all 12 k-mers is presented in Table 2.6.

Table 2.6: Library coverage for k-mers of varying length. Note that the exact coverage for 5-mers is 99.99990625 %, thus only three out of 3.2 million 5-mers were not observed.

K-mer Length	Percentage (%)
1	100
2	100
3	100
4	100
5	100.0
6	96.0
7	25.0
8	1.3
9	5.3E-02
10	2.0E-03
11	6.6E-05
12	1.6E-06

As shown in Table 2.6, 5-mers were the longest k-mers that still had virtually complete library coverage. While 6-mers could still be useful for analysis, they only had approximately 96% coverage. To increase coverage of the 6-mers, we would need to

construct a larger library, conduct experiments with a larger oversampling of the library, or achieve greater NGS sequencing depth. To further illustrate the limitations of using 6-mers, we examined the 5-mer and 6-mer coverage after selection for a single specimen. In this set, there was 87.7% coverage of 5-mers and 15.1% coverage of 6-mers. The maximum possible 6-mer coverage for this specimen was 20.3%. Therefore, since K-TOPE analyzes a single specimen at a time, 6-mers will not have greater than approximately 20% coverage. In summary, to ensure that we avoided sparse datasets we conducted analysis with 5-mers.

Additionally, the length of a k-mer affects expected observations and therefore affects the dynamic range of enrichments. To illustrate, a k-mer with 1 observation and an expected value of 1 would have an enrichment of 1. In contrast, a k-mer with 1 observation and an expected value of 0.1 would have an enrichment of 10. In the case of low expected values, a single observation will lead to a high enrichment. However, a single observation could possibly be due to noise. Ideally, the expected value should be of order one such that a single observation will not lead to a high enrichment. To identify which k-mer length had an expected value of order one, we calculated the expected values for a set of 1.5 million 12-mer peptides with equal amino acid frequencies of 0.05 (Table 2.7). The 5-mers were the only k-mer length with an expected value of order one, with an expected value of around 4. Note that the actual expected value varies with the amino acid frequencies and the number of sequences. Thus, for 6-mers to have an expected value of order one, we would require approximately 10 times more sequences per specimen. Thus, with the current number of sequences per specimen, k-mers of length 5 were optimal for K-TOPE.

Table 2.7: The expected number of sequences for different k-mer lengths. These calculations assumed a total of 1.5 million sequences and equal amino acid frequencies of 0.05. Only 5-mers have an expected value of order 1.

K-mer Length	Expected
1	9.0E+05
2	4.1E+04
3	1.9E+03
4	84
5	3.8
6	0.16
7	7.0E-03
8	2.9E-04
9	1.2E-05
10	4.4E-07
11	1.5E-08
12	3.7E-10

3 Mapping antibody binding using multiplexed epitope substitution analysis

A more complete understanding of antibody-binding epitopes would aid the development of diagnostics, therapeutic antibodies, and vaccines. However, current methods for mapping antibody binding in epitopes require a targeted experimental approach, which limits throughput. To address these limitations, we developed Multiplexed Epitope Substitution Analysis (MESA) which can rapidly characterize multiple epitopes using millions of antibody-binding peptides. We selected peptides from a random 12-mer library that bound to human serum antibody repertoires and determined their sequences using next-generation sequencing (NGS). Next, we evaluated the enrichment of all 5- and 6-mers in the peptide dataset. Computationally, we divided target epitope sequences into overlapping k-mers. Then, the positions in each k-mer were substituted with all 20 amino acids and the enrichments of the substituted k-mers were determined in the peptide dataset. This approach enabled the identification of substitutions favored for binding at each position in the target sequence, revealing the binding motif. To validate MESA, we determined binding motifs for monoclonal antibodies spiked into a serum specimen, recovering the expected binding positions and amino acid preferences. To characterize epitopes bound by a population, we analyzed 50 serum specimens to determine the binding motifs within various target epitopes, including known pathogen epitopes. Binding motifs identified by MESA agreed with those discovered using alternative computational approaches. MESA's ability to utilize the depth of NGS datasets enabled the identification of an Epstein-barr virus binding motif that was not discovered with alternative approaches. These results demonstrate that MESA can rapidly

identify binding motifs for multiple epitopes in parallel to enhance our understanding of antibody interactions.

**This chapter was co-authored with Joel D. Bozekowski with equal contributions.*

3.1 Introduction

The capability to map antibody binding in epitopes has become essential in applications ranging from basic research to therapeutic and diagnostic development. For example, therapeutic antibody development requires precise determination of epitopes to achieve desired biological activity [145] and to avoid undesired cross reactivity [7]. Similarly, epitope identification can aid in the development of vaccines that generate neutralizing antibody responses [6]. Additionally, the performance of many antibody serology diagnostic tests is limited by knowledge of the most sensitive and specific epitopes [34]. Finally, the identification of epitopes can yield more effective affinity-capture reagents for research [8].

While the gold standard for characterizing antigen epitopes is X-ray crystallography [146], epitope mapping methods using substitution analysis and peptide libraries have become commonplace [2,19]. To determine the extent that each position in an epitope contributes to binding, alanine scanning mutagenesis has commonly been employed [28]. Extending this approach, exhaustive mutagenesis can be used wherein each position in an epitope is mutated to all amino acids [26]. Given the labor-intensive nature of these methods, there remains a need for more efficient, high-throughput methods to characterize and map antibody binding in epitopes.

Peptide microarrays have been used extensively to determine antibody specificities [44]. In these approaches, the antigen is tiled into overlapping peptides and exhaustively

mutated. The importance of each amino acid for antibody binding in the entire protein sequence can then be inferred from the extent of antibody binding to each peptide. Unfortunately, since this method uses a targeted library, new microarrays must be prepared for each protein of interest. For the analysis of many antigens, or antibody specificities within unknown antigens, this process becomes impractical.

To address the limitations of targeted epitope characterization approaches, we developed a method termed Multiplexed Epitope Substitution Analysis (MESA), which utilizes random peptide libraries, NGS, and bioinformatics to simulate exhaustive mutagenesis of arbitrary epitopes. First, we selected antibody-binding peptides from a large surface-displayed peptide library. Next, we evaluated the enrichment of all 5-mers or 6-mers in the antibody-binding peptide dataset. We then divided target epitope sequences into overlapping k-mers and evaluated the enrichments of the k-mers and all possible single-amino acid substitutions. Through this analysis, we determined the effect of amino acid substitution at each position in the epitope to reveal the binding positions and amino acid preferences. Since MESA utilizes millions of peptides selected from a random library, many protein epitopes can be characterized simultaneously.

3.2 Results

3.2.1 MESA maps binding in epitopes using random peptide libraries

MESA characterizes the binding of antibodies to linear targets through random peptide library screening. This approach determines the binding motif of a target epitope, indicating which positions are conserved or variable and which amino acids are preferred at each position. We adapted the algorithm ArrayPitope [44] to use peptide sequences derived

from random peptide libraries, rather than microarray binding data. By using random libraries, MESA can examine numerous target sequences via substitution analysis to determine binding motifs (Figure 3.1). A binding motif can be visually represented as an "epitope logo" that reveals amino acid preferences at each position in a target sequence. For this approach, large peptide sequence datasets are obtained by enriching a random peptide library for serum antibody binding and identifying the enriched peptides using next-generation sequencing (NGS) [43]. The epitope region of interest (the target sequence) is first transformed into k-mer sequences (5- or 6-mers) with one amino acid overlap spanning the entire sequence. Then, the enrichment of each k-mer in the antibody-binding peptide dataset is calculated. Each position of the overlapping k-mers is substituted with each amino acid and the enrichments of the substituted k-mers are calculated to determine the effect of amino acid substitution. Finally, the effects of substitution on each k-mer are compiled to determine the effects of substitution in the whole target sequence.

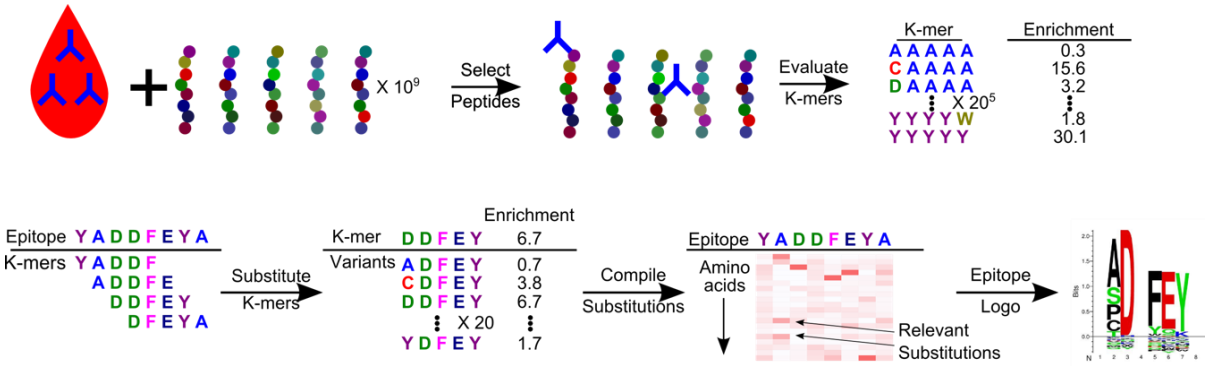


Figure 3.1: An overview of Multiplexed Epitope Substitution Analysis (MESA). A random peptide library is screened for peptides that bind serum antibodies and the enrichments for all k-mers of a set length are calculated. A target epitope is divided into k-mers with a single amino acid overlap. For each k-mer, the positions are substituted with all amino acids to generate 20 variants. Enrichments for all k-mers are then compiled to determine statistically significant positions and valid substitutions in the epitope. The amino acid preferences for each position in the epitope are displayed in an epitope logo. Positions which are not important for binding are blank.

3.2.2 Determining binding motifs for monoclonal antibodies with known epitopes

To validate MESA, we analyzed antibody species with known linear epitopes to determine binding motifs. First, two mAbs, anti-cMyc and anti-HA, were spiked into a human serum specimen at 200 nM each. The known linear epitopes for anti-cMyc and anti-HA are EQKLISEEDL and YPYDVPDYA, respectively. We identified 619,527 12-mer antibody-binding sequences by screening this specimen. MESA was applied to each linear mAb epitope by dividing the epitopes into 5-mers and evaluating 5-mer enrichments in the peptide dataset. The results generated by MESA are visually displayed with alignment heatmaps (Figure 3.2A,C) and epitope logos (Figure 3.2B,D). The relative frequencies of amino acids at each position in the target sequence are displayed in an epitope logo. In an epitope logo, the total height of letters at a position represents the position's importance to binding and the heights of individual letters reflects the amino acid preferences. Blank positions indicate the position was not statistically significant. MESA also generates a regular expression that represents the epitope logo. For each epitope, alignment heatmaps were generated to visualize the importance of each position within the k-mers. In an alignment heatmap, the substituted and target k-mer enrichments are summed for each position in a k-mer. Lower values represent a greater effect of substitution at a position because at an insignificant position, all substituted k-mers will have the same enrichment, leading to a 20 times higher enrichment total than for a significant position. Additionally, the contribution of each k-mer to the epitope logo is proportional to its total enrichment, which corresponds to the total of each row.

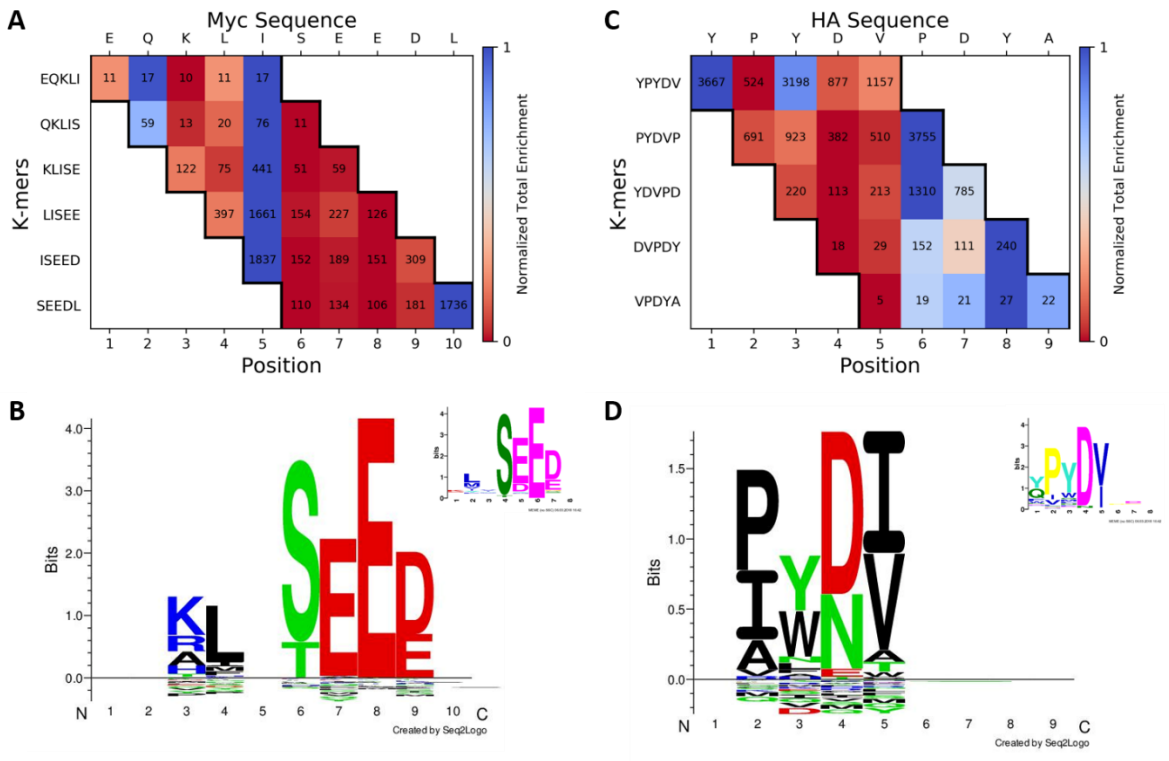


Figure 3.2: MESA determined binding motifs for mAbs. A random bacterial display peptide library was screened against human serum spiked with anti-cMyc and anti-HA mAbs, each at 200 nM, resulting in 619,527 12-mer sequences. MESA was applied to this dataset to generate an (A) alignment heatmap and (B) epitope logo for the cMyc target EQKLISEEDL, and an (C) alignment heatmap and (D) epitope logo for the HA target YPYDVPDYA. MESA used 5-mers, a 10% score threshold, and a 25% minimum enrichment threshold. In alignment heatmaps, the substituted and target k-mer enrichments are summed for each position in the k-mers. Lower values represent a greater effect of substitution at a position. For epitope logos, the absolute height of letters represents the relative effect of substitution at a position. The height of individual letters reflects the binding preference at that position relative to the original k-mer. MEME sequence logos (insets) were obtained via MEME analysis of 5,000 sequences using libraries screened with each mAb.

For the two mAbs tested, MESA identified distinct amino acid preferences at each position that were significant for binding. For example, position 8 of the cMyc epitope logo (Figure 3.2B) was highly conserved for binding with almost exclusive preference for glutamic acid (E), while position 9 was half as conserved with roughly equal preferences for aspartic acid (D) and glutamic acid (E). The regular expressions corresponding to the binding

motifs were KLxSEE[DE] and [PI]Y[DN][IV] for cMyc and HA, respectively. Repeating the analysis with 6-mers showed that epitope logos were sensitive to the k-mer length chosen (Chapter 3.5.1). To validate these epitope logos, we determined mAb binding motifs by screening each mAb spiked into buffer at 20 nM. After NGS, the 5,000 peptides with the highest observations from each library were analyzed using MEME to identify sequence logos for each mAb (Figure 3.2B,D insets). While MEME and MESA identified similar binding motifs for mAbs spiked into buffer, only MESA could identify binding motifs for mAbs spiked into serum. Thus, MESA precisely determined the mAb binding motifs in the presence of background serum antibodies.

3.2.3 Using MESA to identify binding motifs with a single serum specimen

When exact antibody targets are unknown, MESA can instead be used with single peptides that were enriched for antibody binding. MESA can then reveal epitopes within the enriched peptides. We analyzed a single serum specimen, which had 364,411 antibody-binding peptides. Epitope logos and alignment heatmaps were generated for two of the most observed peptide sequences in the library, YADVFEYQYDWP (P1) and TWRDWWSKQPFQ (P2) with 1,429 and 611 observations, respectively (Figure 3.3). For P1, the regular expression was ADxFEY which, along with the alignment heatmap and epitope logo, indicated strong amino acid preferences at all positions except position 3 (Figure 3.3A,B). MEME analysis of the 5,000 highest enriched antibody-binding peptides revealed a highly similar motif (Figure 3.3B inset), suggesting that MESA identified the true binding motif for the target peptide P1. Similar success was obtained for P2, which had a regular expression of S[WF][KR]xW[FYW] and an epitope logo that was nearly identical to the MEME sequence logo (Figure 3.3C,D). Notably, even though the P2 target peptide

TWRDWWSKQPFQ contained a tryptophan (W) at position 6, MESA accurately identified the preference for phenylalanine (F) and tyrosine (Y) as well.

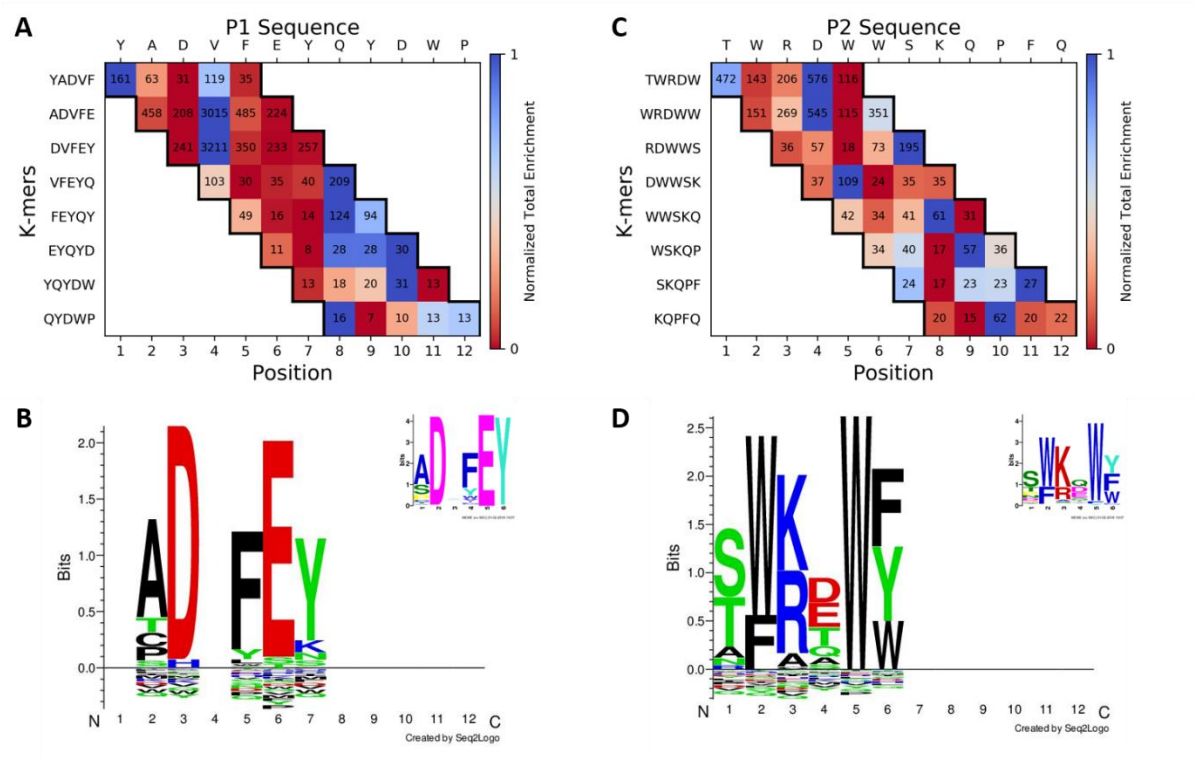


Figure 3.3: MESA determined binding motifs for antibodies in an individual serum specimen. We analyzed two highly enriched antibody-binding peptides using MESA with a single specimen dataset containing 364,411 sequences. MESA generated an (A) alignment heatmap and (B) epitope logo for the P1 target YADVFEYQYDWP, and an (C) alignment heatmap and (D) epitope logo for the P2 target TWRDWWSKQPFQ. MESA used 5-mers, a 10% score threshold, and a 25% minimum enrichment threshold. MEME sequence logos (insets) were obtained via MEME analysis of 5,000 sequences obtained from the specimen library.

3.2.4 Identifying binding motifs using multiple serum specimens

By using large peptide datasets from multiple specimens, MESA identified binding motifs that represented a population. Peptide libraries screened against eight individual serum specimens were sequenced to obtain a total of $1x10^7$ sequences. This dataset was sufficiently large that MESA could utilize 6-mers for increased resolution relative to 5-mers. Also, this analysis used increased parameter stringencies due to the large sequence dataset. We used

MESA to analyze the two peptides with the most observations in at least six specimen libraries, DPYLPHWSTVEV (P3) and KYAFPQRIFVSS (P4) (Figure 3.4). For P3, MESA identified the regular expression as LPHW, with highly conserved residues at positions 4–7 (Figure 3.4A,B). The epitope logo for P3 agreed with the MEME sequence logo determined using all 1,865 sequences present in at least six of the eight libraries. For P4, the regular expression was KxxFPQx[IV], in strong agreement with the MEME sequence logo (Figure 3.4C,D). These results show that MESA accurately characterized the binding of two antibody species present in multiple serum specimens.

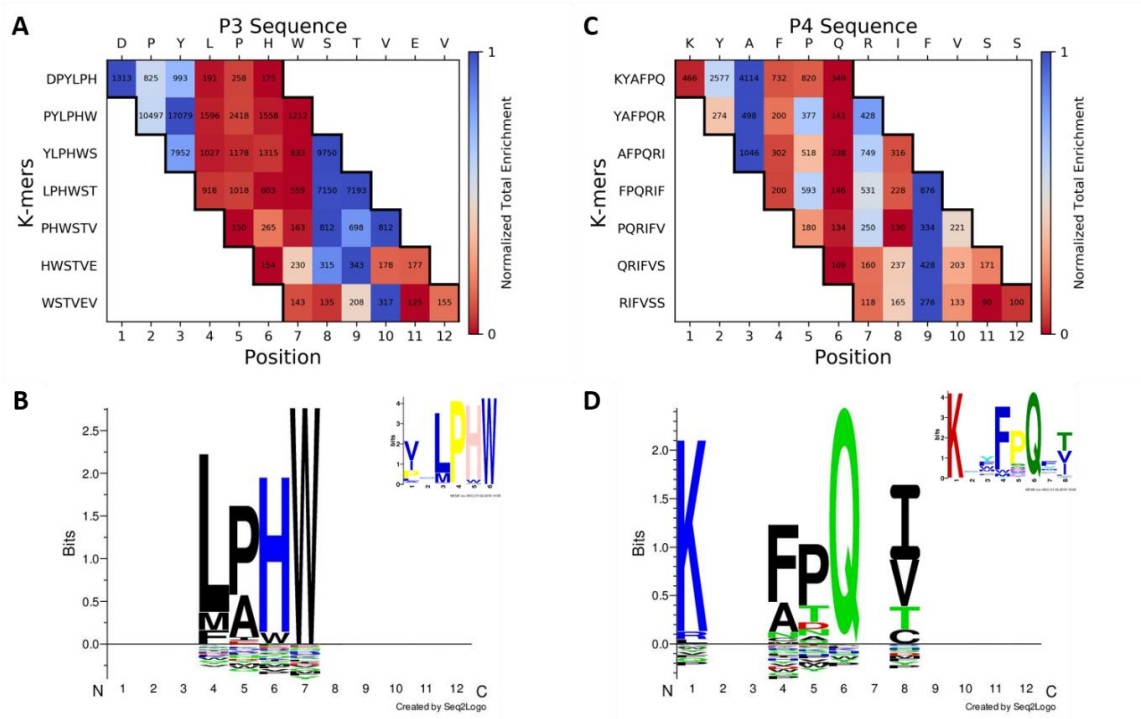


Figure 3.4: MESA identified binding motifs that were common in multiple specimens. Libraries screened against eight individual serum specimens were sequenced to obtain a total of $1x10^7$ 12-mer sequences. The two peptides with the most observations in at least six specimen libraries, DPYLPHWSTVEV (P3) and KYAFPQRIFVSS (P4), were used as targets for MESA. MESA generated an (A) alignment heatmap and (B) epitope logo for the P3 target DPYLPHWSTVEV, and an (C) alignment heatmap and (D) epitope logo for the P4 target KYAFPQRIFVSS. Due to the larger sequence dataset in this analysis, MESA used 6-mers, a 5% score threshold, and a 50% minimum enrichment threshold. To validate the MESA binding motifs, we identified motifs using MEME analysis of all 1,865 sequences observed in at least six of the eight libraries (insets).

To validate MESA with an even larger population of specimens, we analyzed epitope sequences from common antigens using 50 specimens (> $1x10^8$ sequences). We utilized MESA to generate an epitope logo for the common epitope SGSPPRRPPPGRRPFFHPVG from Epstein-Barr virus nuclear antigen 1 (EBNA1) [23] (Figure 3.5A). The regular expression generated for this epitope was RRP[FW]FHP, which was highly enriched in 64% of the specimen libraries. Here, "highly enriched" signifies that a regular expression had an enrichment > 10 in a specimen's sequence set. The EBNA1 motif RRPFF has been found in multiple previous analyses [21,128]. The binding motif for another EBNA1 epitope, EADYFEYHQEGGPDGEPDVP [128], was determined using MESA (Figure 3.5B). The regular expression for this epitope was ADYxEY, which is the same specificity identified with MESA using P1 (Figure 3.3). This motif was highly enriched in 30% of the specimen libraries. The third epitope analyzed, VPEFKGSLP, was from extracellular matrix protein-binding protein emp in *Staphylococcus aureus* [39]. MESA generated the binding motif for this epitope and identified the regular expression VPEFxG[AS], which was highly enriched in 92% of libraries (Figure 3.5C). As additional validation, we conducted a MEME analysis of 5,000 sequences observed in at least 10 of the 50 specimen libraries. This analysis revealed motifs that corresponded to the binding motifs determined with MESA (Figure 3.5 insets). These analyses demonstrated the power of MESA to efficiently mine large datasets for antibody binding characterization.

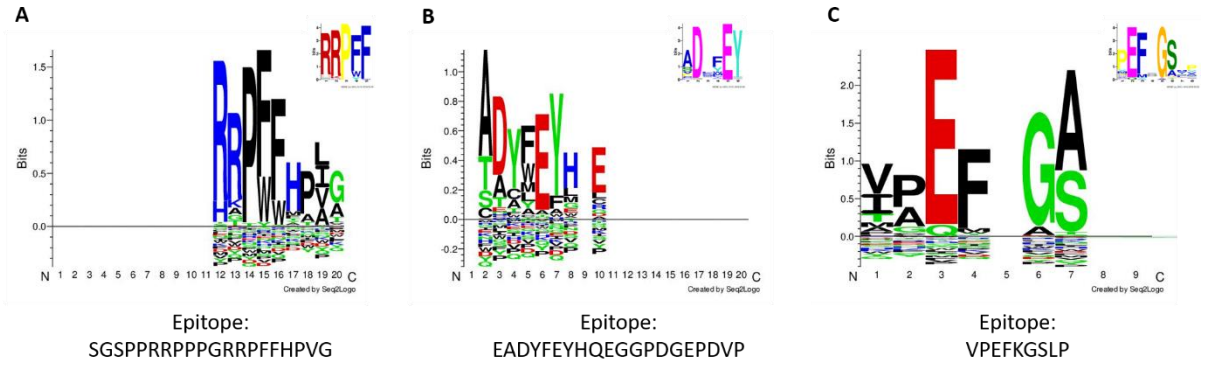


Figure 3.5: Binding motifs of common epitopes were identified using MESA. Libraries from 50 specimens (> $1x10^8$ sequences) were analyzed with MESA to determine the binding motifs of three common epitopes: (A) the EBNA1 epitope SGSPPRRPPPGRRPFFHPVG, (B) the EBNA1 epitope EADYFEYHQEGGPDGEPDVP, and (C) the extracellular matrix protein-binding protein emp epitope VPEFKGSLP from *Staphylococcus aureus*. MESA utilized 6-mers, a 2.5% score threshold, and a 50% minimum enrichment threshold for (A) and (B), but a 25% enrichment threshold was used for (C) due to the shorter epitope length. MEME sequence logos (insets) were obtained via MEME analysis of 5,000 sequences observed in at least 10 of 50 specimen libraries.

Importantly, MESA can generate binding motifs for epitopes that are represented in a small fraction of the sequences, and would therefore be difficult to discover using algorithms like MEME. We utilized MESA to determine the binding motif for another common EBNA1 epitope, RPQKRPSCIGCKGTHGGTGA [23] (Figure 3.6). We determined the regular expression for the binding motif as CIGCR, but we did not identify a corresponding motif using MEME. However, we determined CIGCR to be highly enriched in 34% of specimen libraries, suggesting that this epitope was bound by a significant proportion of the population.

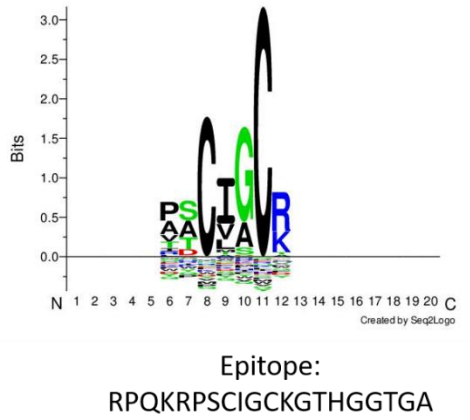


Figure 3.6: An EBNA1 binding motif was discovered using MESA. Libraries from 50 specimens (> $1x10^8$ sequences) were analyzed with MESA to discover a binding motif for the EBNA1 epitope RPQKRPSCIGCKGTHGGTGA. MESA utilized 6-mers, a 2.5% score threshold, and a 50% minimum enrichment threshold. MEME was not able to identify a corresponding motif due to the limited number of input sequences.

3.3 Discussion

Here, we introduced an algorithm for determining the most important positions and amino acid preferences in epitopes, starting from a random peptide library. Notably, this method enables the identification of binding motifs for multiple epitopes, avoiding the need to bias experiments. MESA computationally substitutes individual positions in an epitope and calculates the effects of the substitutions. Positions which are the most affected by substitutions are plausibly important to antibody binding. Since our bacterial display library has ~100% coverage of 5-mers and ~96% coverage of 6-mers, our approach can utilize k-mers to precisely evaluate the effect of a substitution in a peptide. It may be difficult to correctly identify binding motifs for approaches with smaller libraries, such as microarrays with low k-mer coverage [35]. By analyzing known epitopes for mAbs spiked into serum, we showed that only a few positions dominated binding for these antibodies. Additionally, by analyzing antibody-binding peptides and known pathogen epitopes, we could refine epitopes

into shorter motifs representing the most important binding positions. We mainly focused on common epitopes from the antigen EBNA1, since Epstein-Barr virus infects over 90% of the population [147]. We were able to validate our results by comparison to MEME [49], with the exception of the binding motif CIGCR (Figure 3.6). However, the binding motif CIGCR was likely genuine, since it was enriched in the specimens' sequences and has been previously reported [42]. A limitation of MEME is that it can only analyze 5,000 sequences, which for the single specimen and 50 specimen analyses was 1% and 0.005% of all sequences, respectively. Therefore, an advantage of MESA is that it can more thoroughly explore the depth of NGS datasets. Another limitation of using MEME to find significant positions in a target sequence is that it required the cumbersome process of searching through motifs for similarity to the target. In contrast, MESA simply substitutes a target sequence and returns the binding motif.

MESA has significant advantages over other epitope substitution approaches. Targeted substitution schemes like ArrayPitope tile peptides from a protein sequence in a microarray format, and then mutate individual positions to all other amino acids [44]. The main issue with using a non-random library is that experiments must be repeated for each additional protein target. While researchers have used random bacteriophage libraries to identify epitopes [39], these approaches required separate targeted libraries to map binding in epitopes. Non-random libraries can also lead to inaccurate results if serum antibodies have a preference towards a variant of a protein sequence. For example, Celiac disease antibodies target deamidated gliadin, thus a microarray using the gliadin sequence could miss this binding interaction [148]. MESA represents the first algorithm that can multiplex epitope substitution using random libraries. With random libraries, there is no need to bias the

experimental approach towards a specific antigen, thus, epitopes corresponding to many antigens can be analyzed.

Limitations of MESA are mostly related to its inclusion of noise, a necessary result of using large NGS datasets. With large datasets, sequences which are irrelevant to the true binding motif are often included, leading to a poor signal-to-noise ratio. Another limitation is that this method requires $> 10^5$ sequences for the analysis of serum antibodies, though for typical NGS datasets, this requirement should not be difficult to achieve [43]. Also, if a group of subjects does not have homogenous binding to an epitope, MESA may not generate a clear binding motif.

MESA could be used to map antibody binding in epitopes for the development of more effective diagnostics, therapeutics, and vaccines. Generating a binding motif for a disease-specific epitope would allow for the optimization of diagnostic peptides [148]. Additionally, associating motifs with disease antigens could provide insights into etiology [42,149] and lead to the development of antibody therapeutics, such as for cancer [150]. MESA could aid the study of broadly neutralizing antibodies for vaccine design and therapeutic development in infections such as HIV [151,152]. This approach could also be used to improve the development of peptides vaccines [153] enabling a more focused immune response. Moreover, there is an increasing need for robust methods capable of studying and resolving cross-reactivity in vaccine development [154].

MESA can characterize binding motifs for antibody-binding peptides or known protein epitopes. Importantly, MESA has the capability to probe the full depth of large NGS datasets. The ability to identify epitopes relating to multiple antigens will likely become indispensable to fully interrogating antibody repertoires.

3.4 Materials and methods

3.4.1 Bacterial display and sequencing

The protocol for screening bacterial display peptide libraries was carried out as previously described [43,124]. Briefly, we added 1:100 diluted human serum to an *E. coli* display library of $8x10^9$ random 12-mer peptides, and sorted cells with bound antibodies through two rounds of magnetic selection using Protein A/G magnetic beads (Thermo Scientific Pierce). DNA amplicon libraries were prepared from the enriched library cells for sequencing using the Illumina NextSeq. All sera (N=60) were obtained as de-identified specimens from biobanks according to institutional guidelines, (UCSB biosafety authorization numbers #201417, #201713), and handled according to CDC-recommended BSL-2 guidelines.

3.4.2 Monoclonal antibody spike-in

Two rabbit monoclonal antibodies (mAbs), cMyc-Tag 71D10 (EQKLISEEDL) and HA-Tag C29F4 (YPYDVPDYA) (Cell Signaling Technology), were used for MESA analysis of known target sequences. The mAbs were either added to a human serum specimen at 200 nM or added to 1x PBST at 20 nM [140].

3.4.3 Sequence processing

We generated non-redundant sequences from FASTQ files and calculated 5-mer and 6-mer enrichments using an adapted version of IMUNE [43]. Enrichment is defined as the ratio of observations of a k-mer to the “expected” observations. The number of “expected” observations is calculated by multiplying the number of frames the k-mer could fit in a 12-mer peptide with the total number of sequences and the probability of the k-mer appearing

based on amino acid usage. For computations with large specimen datasets using 6-mers, we sum the enrichments from all specimens for each k-mer. Running these programs requires at least 16 GB of free RAM and 100 GB of hard-drive space. Analysis was carried out on a Dell Optiplex 9020 with an Intel® Core™ i7-4790 CPU @ 3.60 GHz, 64-bit operating system, and 32.0 GB of RAM. Processing FASTQ files into subsequences from 12 specimens (~1.5 million unique sequences per specimen) and calculating 5-mer and 6-mer enrichments required 136 minutes, 10.1 minutes, and 140 minutes, respectively. The duration of these calculations scales approximately linearly with the number of specimens and sequences.

3.4.4 MESA algorithm

Multiplexed epitope substitution analysis (MESA) determines binding motifs, which show the positions in an antibody-binding peptide that are important for binding and the amino acid (AA) preferences at each binding position. A binding motif is displayed as an epitope logo in which the effect of AA substitution at a position corresponds to the absolute height at that position. Positions with an insignificant contribution to binding have a height of zero. The height of individual AA letters indicates the relative importance of an AA at a position.

The approach for determining binding motifs is based on the methodology used in ArrayPitope [44]. To start, a sequence with known antibody binding, denoted the target sequence, is tiled into k-mers of 5 or 6 AA with a single AA overlap. For each k-mer, one position at a time is substituted with all AAs. This results in 100 variants for 5-mers and 120 variants for 6-mers. The k-mer which corresponds to the native antigen sequence is termed the original k-mer. When using multiple specimens, the enrichment of a k-mer is the sum of enrichments in all specimens. All the enrichments are normalized by the original k-mer

enrichment. The normalized enrichments are used to populate a position-specific scoring matrix (PSSM). In a PSSM, columns represent positions in the target sequence and rows represent AA substitutions.

The average of each column in the PSSM, denoted the score, indicates whether a position in a k-mer is conserved ('N') or variable ('X'). The "conservation string" for a k-mer indicates which positions are conserved (e.g. NXNXN). If a conserved position is substituted, antibody binding will diminish, whereas antibody binding is nearly unaffected by the substitution of a variable position. A score near 1 indicates that a position has little preference for the original AA at that position and is it therefore a variable position. If the score is far below 1, this indicates a strong preference for the target sequence AA and it is considered a conserved position. Therefore, we use a binary cutoff value on the score, termed the score threshold, to determine if a position should be considered conserved or variable. The score threshold is determined by generating PSSMs for 1,000 random k-mers and compiling their scores into a list. The assumption is that these random k-mers should have all variable positions and should thus possess scores approximating random chance. A score threshold is chosen using a low percentile (e.g. 5%) of the list of random scores. Additionally, a position is considered variable if the sum of enrichments for the 20 variants that substitute that position ("the enrichment sum") is less than the minimum enrichment threshold. This threshold excludes spurious results where a position appears conserved because the enrichment is too low. The minimum enrichment threshold is defined as a percentile in the distribution of all enrichment sums generated from a target sequence. To determine a "sequence conservation string", the conservation string for each k-mer is aligned

to the target sequence. If there is at least one k-mer that is conserved at a position in the alignment, then that position is conserved in the sequence conservation string.

To determine the binding motif, a substitution matrix is generated for each position in the target sequence. A substitution matrix describes the AA preferences for each position (the "substitution position") in the target sequence. PSSM columns from positions that overlap with the substitution position are used to construct a substitution matrix. PSSM columns are only included in the substitution matrix if they are conserved in the k-mer. To determine the frequency of each AA at a substitution position, we determine the average of each column in the substitution matrix and normalize the averages by the sum of all column averages.

From the substitution matrices, each position in the target sequence has AA frequency values. The frequency values are used to generate the relative entropy matrix by applying the following formula to each AA frequency:

$$entropy = f_{aa} \log_2 \frac{f_{aa}}{u_{aa}} \tag{3.1}$$

where f_{aa} is the frequency of the AA and u_{aa} is the usage of the AA in the peptide dataset. In the relative entropy matrix, each row corresponds to a position in the initial sequence and each column corresponds to one of the 20 AAs. If a position is determined to be variable, it is represented by a row of zeros so that it will not appear in the epitope logo. The relative entropy matrix is then input into Seq2Logo to generate an epitope logo [155]. Generating the relative entropy matrix for the peptide YADVFEYQYDWP (P1) required 0.016 seconds. Generating the epitope logo (using Seq2Logo) from the relative entropy matrix required 0.75 seconds.

Determining a regular expression allows for a single textual representation of the binding motif. To determine the regular expression, each position in the target peptide

becomes an ‘X’ if it is variable or is replaced with one or more AAs if the position is conserved. AAs are included in the regular expression if they meet a frequency cutoff (0.2 in this analysis). Leading and trailing ‘X’s are then trimmed from the regular expression.

Elements that were adapted from ArrayPitope [44] include dividing a target sequence into shorter overlapping subsequences, generating PSSMs and substitution matrices, and visualizing the results using Seq2Logo. MESA differs from ArrayPitope by using peptides selected from a random library rather than from targeted microarrays, dividing sequences into k-mers, using a different statistical approach, and generating an entropy matrix for input to Seq2Logo.

3.4.5 Identifying antibody binding motifs with MEME

To validate the binding motifs discovered using MESA, the motif discovery algorithm MEME [49] was used to determine antibody binding specificities in a set of peptide sequences. MEME uses pairwise sequence comparisons in small sequence sets of less than about 5,000 members, while MESA utilizes substitution analysis throughout full NGS datasets. Although variations in the results from MESA and MEME will exist due to differences in algorithms and data input, comparison to MEME is effective for broadly confirming antibody binding specificities and assessing MESA performance.

MEME is not suitable for analyzing large peptide datasets due to run-time constraints. Therefore, all MEME analyses were run with a maximum of 5,000 sequences. To identify mAb motifs, a random peptide display library was screened against each mAb at 20 nM in 1x PBST [140] and sequenced. A minimum motif width of 8 was utilized for the MEME mAb analyses. For all analyses of serum antibody motifs, a minimum motif width of 4 was used. For 5,000 sequences, MEME analysis required 10.0 ± 1.4 hours.

3.5 Supplemental analysis of MESA

3.5.1 Effects of MESA parameter selection on binding motifs

We examined the effects of varying MESA parameters on binding motifs by examining the changes in epitope logos. First, we varied the score threshold, which determines whether a given position in a k-mer is conserved. We completed this analysis using both 5-mers and 6-mers for two mAbs (Figure 3.7). Whereas epitope logos generated using 5-mers changed smoothly as score threshold was increased, epitope logos generated using 6-mers changed dramatically when slightly varying the score threshold. This suggested that that the mAb datasets contained too few sequences for 6-mer analysis. As a result, we concluded that for smaller peptide datasets, MESA should be restricted to analysis using 5-mers.

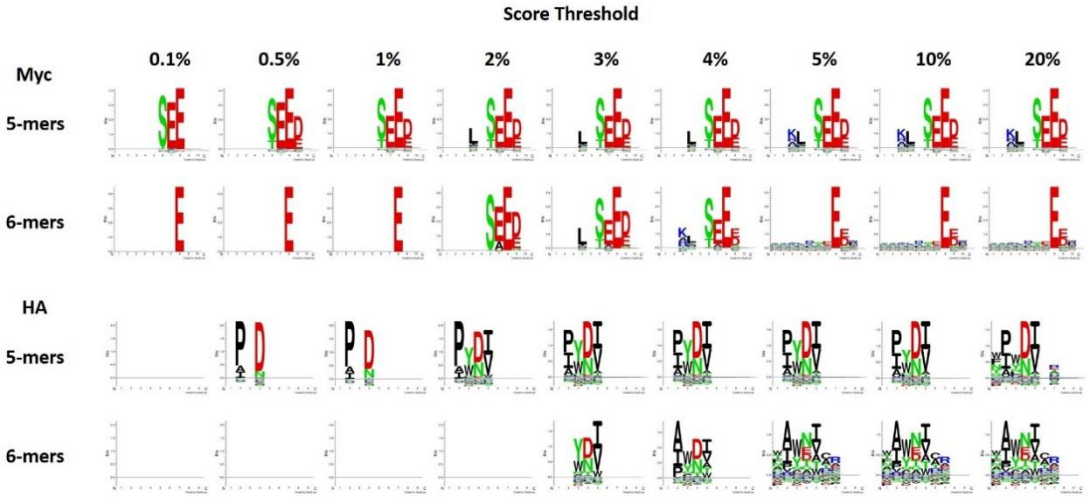


Figure 3.7: Epitope logos generated for monoclonal antibodies with varying score thresholds. Epitope logos for anti-cMyc (EQKLISEEDL) and anti-HA (YPYDVPDYA) mAbs were determined with MESA using 5-mers and 6-mers. The score threshold was varied to observe the effect on epitope logos. MESA was used with a 25% minimum enrichment threshold for all analyses. For MESA with 6-mers, the score threshold had a significant impact on the epitope logo due to scarcity of 6-mers in this small dataset. For 5-mers, the epitope logos were largely constant over a wide range of score thresholds.

Another parameter which could affect epitope logos is the minimum enrichment threshold required for a position to be deemed statistically significant. If the total enrichment at a position does not exceed this threshold, it will be considered a variable position. This threshold largely controls for false positive positions, which have low scores (higher significance) simply due to the sparsity of k-mers. Epitope logos were generated for the mAb datasets using 5-mers while varying the minimum enrichment threshold (Figure 3.8). However, the epitope logos were minimally affected unless the minimum enrichment threshold was raised above 50%. To observe the effect of the minimum enrichment threshold on a larger dataset using 6-mers, we varied the minimum enrichment threshold for the P3 and P4 targets from the multiple specimen analysis (n=8) (Figure 3.9). For this larger dataset, varying the minimum enrichment threshold had a significant effect on the epitope logos. Thus, parameter selection was highly influenced by dataset size.

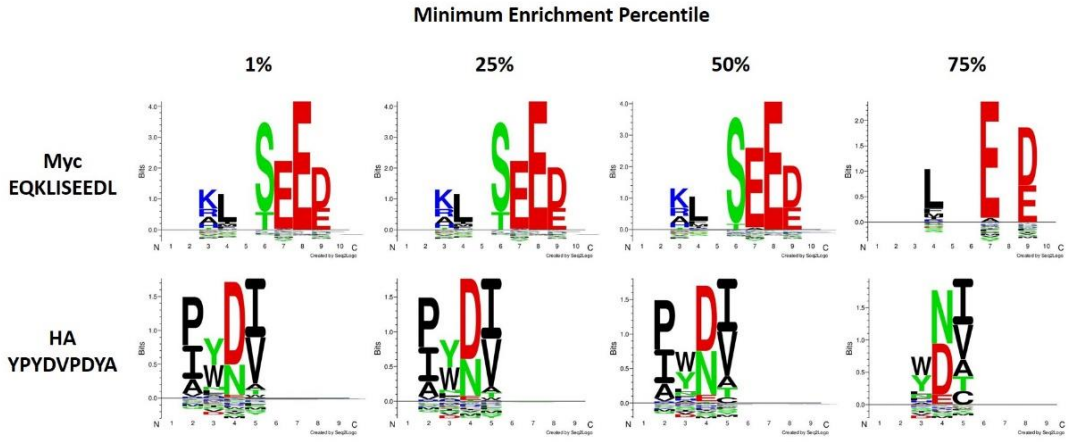


Figure 3.8: Monoclonal antibody epitope logos generated with varying minimum enrichment threshold percentiles. MESA epitope logos for anti-cMyc (EQKLISEEDL) and anti-HA (YPYDVPDYA) mAbs were determined using 5-mers while varying the minimum enrichment threshold. The minimum enrichment threshold did not have a large impact on the epitope logos. A 10% score threshold was used for all analyses.

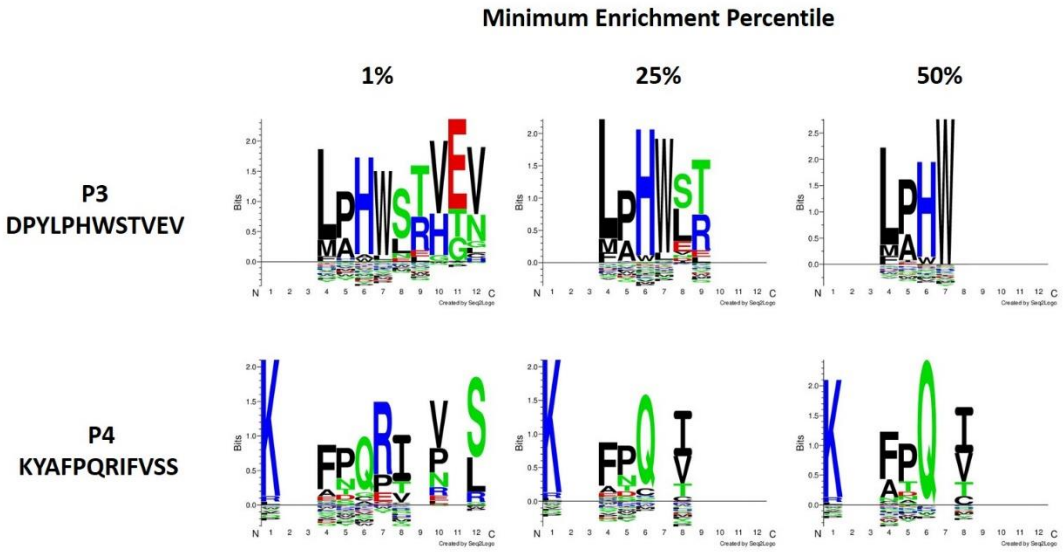


Figure 3.9: Epitope logos were generated for multiple specimens with varying minimum enrichment threshold percentiles. MESA epitope logos were determined for targets P3 (DPYLPHWSTVEV) and P4 (KYAFPQRIFVSS) from eight specimen libraries ($1x10^7$ sequences) using 6-mers while varying the minimum enrichment threshold. A 5% score threshold was used for all analyses.

*The following section was **not** co-authored with Joel D. Bozekowski.

3.5.2 Determining binding motifs for K-TOPE epitopes

MESA was used to determine binding motifs for the epitopes identified by K-TOPE in Chapter 2. Binding motifs were first generated for epitopes bound by the antibodies against cMyc, V5, and amyloid beta using a 10% score threshold (Figure 3.10). For V5 and amyloid beta, the binding motifs contained 5 positions with clear amino acid preferences. However, the binding motif for anti-cMyc did not have a regular expression similar to KLxSEE[DE] identified for anti-cMyc earlier (Figure 3.2). This may be because the cMyc antibody used for K-TOPE analysis was polyclonal (Figure 3.10A), whereas the anti-cMyc used earlier was monoclonal (Figure 3.2).

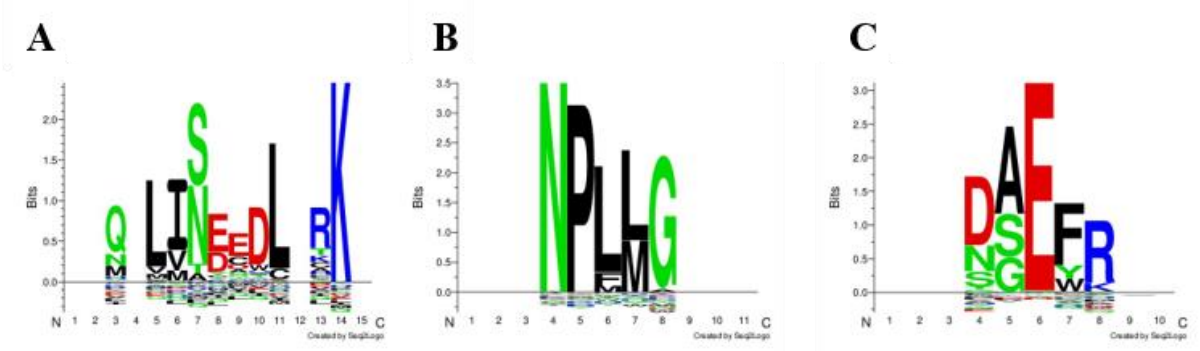


Figure 3.10: Binding motifs identified for 3 antibodies of known specificity. Epitope logos were generated for the epitopes (A) EEQKLISEEDLLRKR from cMyc, (B) VKMDAEFRHD from amyloid beta, and (C) PIPNPLLGLDS from V5. The regular expressions were (A) QXL[IV][SN][ED]EDLXRK, (B) NPL[LM]G, and (C) [DNS][ASG]E[FY]R.

We also used MESA to determine binding motifs for highly prevalent bacterial and viral epitopes (Chapter 2). Due to the large size of the dataset (250 specimens), we used a 2.5% score threshold to reduce noise. We generated epitope logos (Figure 3.11) for four epitopes in the *Rhinovirus A* genome polyprotein (Figure 2.4). The epitope logos showed that these prevalent *Rhinovirus A* epitopes generally had only 4-7 conserved positions, which is consistent with the expectations that approximately 5 positions dominate binding [46]. Additionally, this analysis revealed whether each epitope was bound in the beginning of the sequence (Figure 3.11A), the end of the sequence (Figure 3.11B and Figure 3.11C), or the middle of the sequence (Figure 3.11D).

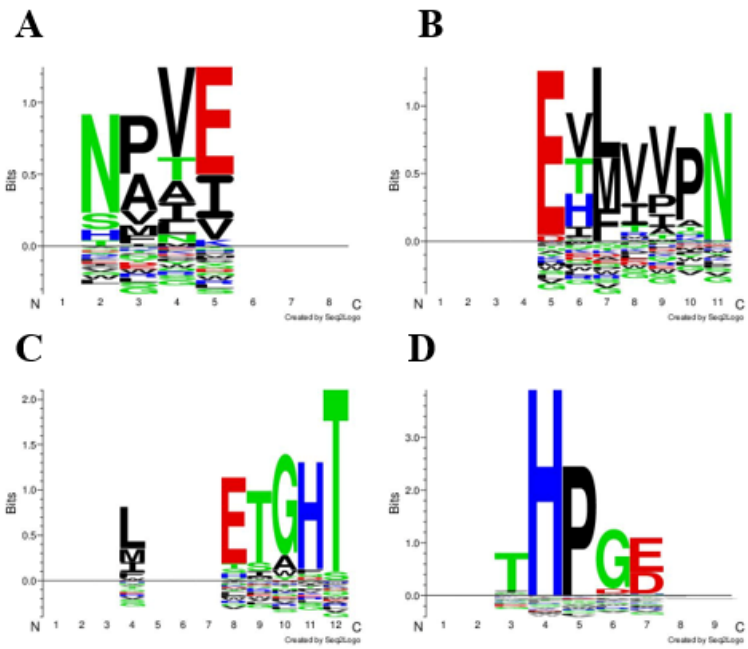


Figure 3.11: Binding motifs identified for four *Rhinovirus A* epitopes. Epitope logos were generated for (A) QNPVENYI, (B) DSVLEVLVVPN, (C) APALDAAETGHT, and (D) NHTHPGEQG. The regular expressions were (A) N[PA]VE, (B), EV[LM]VVPN, (C) LXXXETGHT, and (D) THPG[ED].

We determined binding motifs for three prevalent viral epitopes (Table 2.3) identified by K-TOPE (Figure 3.12). These epitope logos were similar to those presented earlier in the chapter (Figure 3.4B, Figure 3.5A, Figure 3.6). This similarity suggests that determining binding motifs is robust to using different sets of specimens and target epitopes. The epitope logo generated for an EBV epitope (Figure 3.12B) was similar to the epitope logo generated for a highly-observed peptide (Figure 3.4B). Thus, this analysis suggested that the highly-observed peptide contained an EBV epitope.

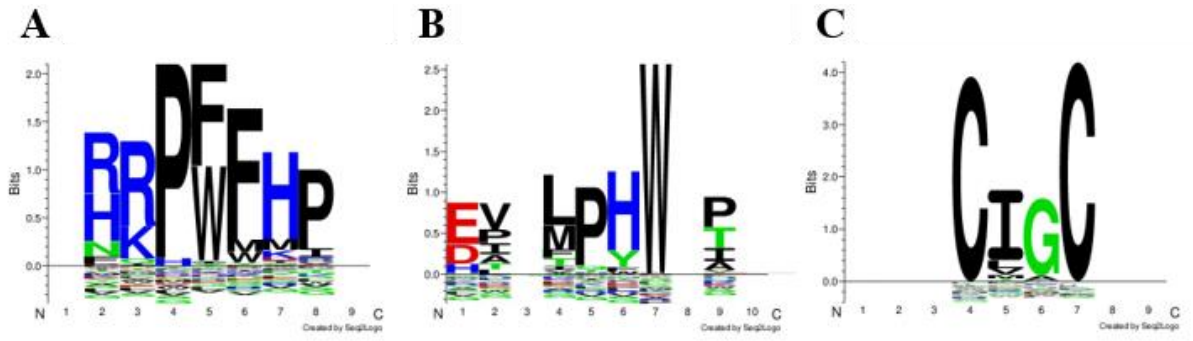


Figure 3.12: Binding motifs identified for three viral epitopes. Epitope logos were generated for (A) GRRPFFHPV, (B) EVKLPHWTPT, and (C) RPSCIGCKG. The regular expressions were (A) [RH][RK]P[WF]FHP, (B), EVX[LM]PHWXP, and (C) CIGC.

Finally, we determined binding motifs (Figure 3.13) for three prevalent bacterial epitopes identified by K-TOPE (Table 2.4). Position 4 was variable in the epitope logo for GQKMDDMLNS (Figure 3.13B) and variable in an alignment of streptolysin O proteins from 7 strains. Thus, these results demonstrated that natural variability in an antigen sequence is reflected in a binding motif. The epitope logos in Figure 3.13A, Figure 3.13C, and Figure 3.5C were similar, yet they were generated from sequences with minimal similarity. Since these epitopes all generated similar binding motifs, it is possible that they are related to a single antibody specificity.

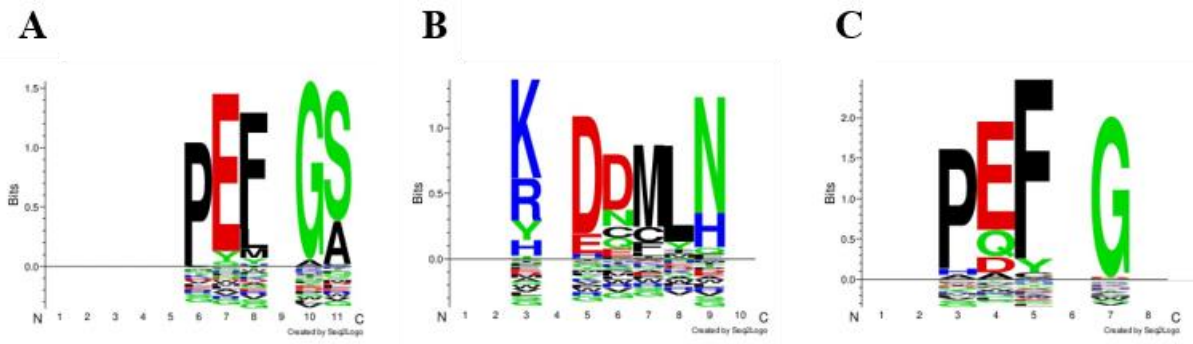


Figure 3.13: Binding motifs identified for three bacterial epitopes. Epitope logos were generated for (A) PTHYVPEFKGS, (B) GQKMDDMLNS, and (C) LIPEFIGR. The regular expressions were (A) PEFXG[SA], (B), [KR]XDDMLN, and (C) P[EQ]FXG.

4 Identification of disease-specific epitope and antigens

In a multitude of diseases, antibodies contribute to and characterize disease progression. Three diseases with prominent antibody responses are age-related macular degeneration, herpes simplex virus, and Chagas disease. Identifying disease-specific epitopes and antigens would facilitate the prevention, diagnosis, and treatment of these diseases. To identify disease-specific epitopes and antigens, we used K-TOPE with disease and control specimens. Then, we mapped antibody binding in these epitopes using MESA. With these tools, we identified 42 epitopes in candidate AMD autoantigens as well as 53 epitopes that tracked with disease progression. Three candidate AMD autoantigens had a clear relation to immune or eye-related pathology. Additionally, we identified 30 HSV2-specific epitopes that were 100% specific against HSV1 in novel and previously reported antigens. Several epitopes were HSV2-specific even though the epitope sequences were in both the HSV1 and HSV2 proteomes. For Chagas disease, we identified 222 epitopes in multiple novel and previously described antigens. Notably, these epitopes exhibited similarity to four out of the seven peptides used in a Chagas-diagnostic assay. We demonstrated that several putative Chagas-specific epitopes cross-reacted with a similar parasite. Also, we identified 1,084 Chagas-specific epitopes in candidate autoantigens. By determining binding motifs, we showed that generally only four or five positions in disease-specific epitopes were important to binding. Applying this approach to additional diseases with antibody responses could help identify medically relevant epitopes and antigens.

4.1 Introduction

Antibody responses play an integral role in the progression of a variety of diseases. The ability to identify and characterize antibody responses would aid the prevention, diagnosis, and treatment of diseases. A disease can be prevented by rationally designing vaccines that elicit potent neutralizing antibodies against a precise epitope [156]. Since treatments are often most effective in the earliest stages of disease, the early detection of disease-related antibodies in patient sera is crucial for effective treatment [30]. Critically, monoclonal antibodies that specifically target an epitope are frequently used therapeutically [157]. Thus, interrogating and understanding the humoral response to diseases can achieve multiple medical aims.

K-TOPE identifies disease-specific epitopes which can then be further characterized using MESA. These tools use antibody-binding peptides to identify binding signatures that differ between disease and control specimens. Examining these differences increases our understanding of disease-specific antibody responses. We used K-TOPE and MESA to analyze age-related macular degeneration (AMD), herpes simplex virus (HSV), and Chagas disease. Applying K-TOPE and MESA to diseases involving human, viral, and parasitic targets constitutes a broad validation of these approaches.

The first disease we analyzed was the degenerative retinal disease AMD, which is the leading cause of blindness in the developed world [57]. AMD is classified as “dry” if it features geographic atrophy or “wet” if it occurs with abnormal blood vessel growth. Anti-retinal antibodies have been identified in AMD patient sera, suggesting that this disease has autoimmune components [71]. While researchers have used protein microarrays to identify autoantigens [80], no study has attempted unbiased discovery of AMD-specific epitopes.

Identifying disease-specific epitopes could aid the discovery of a reliable serological biomarker that could be used for diagnosis [59].

Another disease with antibody responses is the common viral infection HSV, which causes cold sores (HSV1) and genital ulcers (HSV2) [83]. Diagnostic discovery generally focuses on HSV2, since it synergizes with HIV infections [86]. Particularly, researchers have attempted to discover epitopes in glycoprotein G, since it differs substantially between the two HSV species [90]. In general, efforts have been limited to identifying epitopes in the surface-exposed envelope glycoproteins, using approaches such as microarrays [101]. Therefore, it would be novel to identify HSV2-specific epitopes using the entire proteomes of HSV1 and HSV2.

Finally, the protozoan *T. cruzi* is responsible for Chagas disease, which can lead to cardiomyopathy and digestive megasyndromes [102]. Since 30-40% of patients develop cardiac and digestive pathology 10-30 years after initial infection, it is important to diagnose Chagas disease before these symptoms occur [102]. Since *T. cruzi* often has coinfections with the causative agent of leishmaniasis, *L. major*, it is necessary to identify antigens and epitopes that are not bound by leishmaniasis specimens [123]. Also, Chagas disease has autoimmune components [104], which suggests that there could be undiscovered autoantigens. A specific well-characterized Chagas disease antigen is trypomastigote small surface antigen (TSSA) [110], which has been used in a diagnostic test [108]. Additional *T. cruzi* antigens have been identified and characterized using proteome-derived [24] and random [114] libraries. However, there is a lack of studies that have identified autoantigens or have shown that Chagas-specific epitopes do not cross-react with *L. major*.

These diseases have not yet been analyzed using approaches with the flexibility and breadth of K-TOPE and MESA. Therefore, we applied these approaches to identify novel epitopes and antigens.

4.2 Results

4.2.1 Age-related macular degeneration

We analyzed AMD specimens that were originally obtained during the Age-related Eye Disease study (AREDS) [158]. We had 49 specimens that were classified as having either wet AMD, dry AMD, or both types. Of these specimens, 19 were longitudinally collected, such that there were specimens before vision loss (initial timepoint) and during vision loss (final timepoint). Since drusen size is a major predictor of AMD status [65], we had 49 control specimens which had small or medium drusen. The heterogeneity of these specimens posed a challenge for analysis. To allow for sufficiently large sample sizes, we grouped specimens as either disease or control, ignoring any subgroups.

Since autoimmunity is generally accepted to be a component of AMD [71], we analyzed the human proteome for autoantigen epitopes. First, we divided the specimens into a training set (25 disease and 25 control specimens) and a validation set (24 disease and 24 control specimens). All specimens corresponded to initial timepoints to enable the identification of early markers of AMD. To identify epitopes that characterized the full set of AMD specimens, we performed a 2-fold cross validation scheme. In this scheme, we identified 25 epitopes using the training set that were sensitive and specific in the validation set. Then we switched the training and validation sets to identify 28 additional epitopes. Finally, we combined the two sets of epitopes, removed any redundancy, and re-evaluated

the prevalence and specificity using the full set of specimens. Using this approach, we identified 42 epitopes with prevalence > 0.1 and specificity > 0.95 (Table 4.1). Thus, the cross-validation scheme resulted in >50% more epitopes than were identified using separate training and validation sets. Examining the degree to which each specimen bound the epitopes (Figure 4.1) demonstrated that the epitopes were highly specific, though not highly sensitive. It is likely that due to the heterogeneity of the specimen set, only low prevalence epitopes could be identified. None of the antigens matched autoantigens implicated in other autoimmune diseases [159]. However, plausible autoantigens were identified such as Raftlin, which regulates B-cell antigen receptor-mediated signaling [160], and eyes absent homolog 3, which may be involved in the development of the eye [161]. Additionally, Complement C4-A was a plausible antigen since it is involved in the antibody-mediated classical cascade [162], and because complement plays a central role in AMD pathogenesis [163]. Thus, we identified multiple autoantigens with plausible connections to immunological and eye-related pathology.

Table 4.1: AMD-specific epitopes were identified. A total of 42 epitopes with prevalence > 0.1 and specificity > 0.95 were identified using a 2-fold cross validation scheme. Plausible autoantigens are in bold.

Epitope	Protein	Accession	Prevalence	Specificity
ESGALVNFL	Probable small intestine urate exporter	Q9Y2C5	0.184	0.959
ALVNYT	Sideroflexin-2	Q96NB2	0.163	0.959
ALGGLVNAV	Treslin	Q7Z2Z1	0.163	0.959
GGMLVNAV	**Raftlin**	Q14699	0.143	1
PCCFTDLK	N-alpha-acetyltransferase 25, NatB auxiliary subunit	Q14CX7	0.143	0.98
SVWTKTKAA	Protein ERGIC-53	P49257	0.143	0.98
VGNLVNW	SID1 transmembrane family member 1	Q6AXF6	0.143	0.98
GALVND	Probable ribonuclease ZC3H12D	A2A288	0.143	0.98
GAAAAAG	Four-jointed box protein 1	Q8BQB4	0.143	0.959
RQAAAASAAEAG	Putative PIP5K1A and PSMD4-like protein	A2A3N6	0.143	0.959
ASLVNASI	Anillin	Q9NQW6	0.143	0.959
LVNAPY	Phospholipid phosphatase 2	O43688	0.143	0.959
GALVNDE	Transcription initiation factor TFIID subunit 1	P21675	0.122	1
GALVNDE	Isoform 4 of Transcription initiation factor TFIID subunit 1	P21675-4	0.122	1
GELVNAA	Amidophosphoribosyltransferase	P35433	0.122	1
GNLVNF	Fibronectin type III domain-containing protein 8	Q8TC99	0.122	1
SLVNAA	Lysophosphatidic acid receptor 2	Q9HBW0	0.122	1
SKEQHLTF	Ankyrin-3	Q12955	0.122	1
GALVNERTV	Inactive serine protease PAMR1	Q6UXH9	0.122	0.98
TKGLQGTTA	HemK methyltransferase family member 2	Q9Y5N5	0.122	0.98
LKGLQP	Ubinuclein-2	Q6ZU65	0.122	0.98
TMPFDGFDH	NUAK family SNF1-like kinase 1	O60285	0.122	0.98
LKEKHIT	E3 ubiquitin-protein ligase RNF31	Q96EP0	0.122	0.98
KESIFP	Condensin-2 complex subunit D3	P42695	0.122	0.98
AFFHPN	XK-related protein 4	Q5GH76	0.122	0.98
ALVNAI	Ubiquitin-like protein 7	Q96S82	0.122	0.98
TVKEAHLTKD	TBCC domain-containing protein 1	Q9NVR7	0.122	0.98
CKERHFV	2-oxoisovalerate dehydrogenase subunit alpha, mitochondrial	P12694	0.122	0.98
ALVNATE	FACT complex subunit SPT16	Q9Y5B9	0.122	0.98
LVNAQQ	Terminal uridylyltransferase 4	Q5TAX3	0.122	0.98
ALERGLQD	**Complement C4-A**	P0C0L4	0.122	0.98
LKGLQP	Putative ATP-dependent RNA helicase TDRD12	Q587J7	0.122	0.98
LVMGNIIN	SID1 transmembrane family member 2	Q8CIF6	0.122	0.98
GLGPRGLQAT	Laminin subunit alpha-5	O15230	0.122	0.959
TYQSEKPS	**Eyes absent homolog 3**	Q99504	0.122	0.959
GAAKGLQ	Receptor tyrosine-protein kinase erbB-2	P04626	0.122	0.959

FTIFPQF	ATP-binding cassette sub-family A member 13	Q86UQ4	0.122	0.959
LDNKTL	Guanine nucleotide exchange protein SMCR8	Q8TEV9	0.122	0.959
FLNEKR	Protein phosphatase 1 regulatory subunit 26	Q5T8A7	0.122	0.959
SLSRGLQV	Serine/threonine-protein kinase TBK1	Q9UHD2	0.122	0.959
SASAAAASAAAA	REST corepressor 1	Q9UKL0	0.122	0.959
SGVTAEK	PH domain leucine-rich repeat-containing protein phosphatase 1	O60346	0.122	0.959

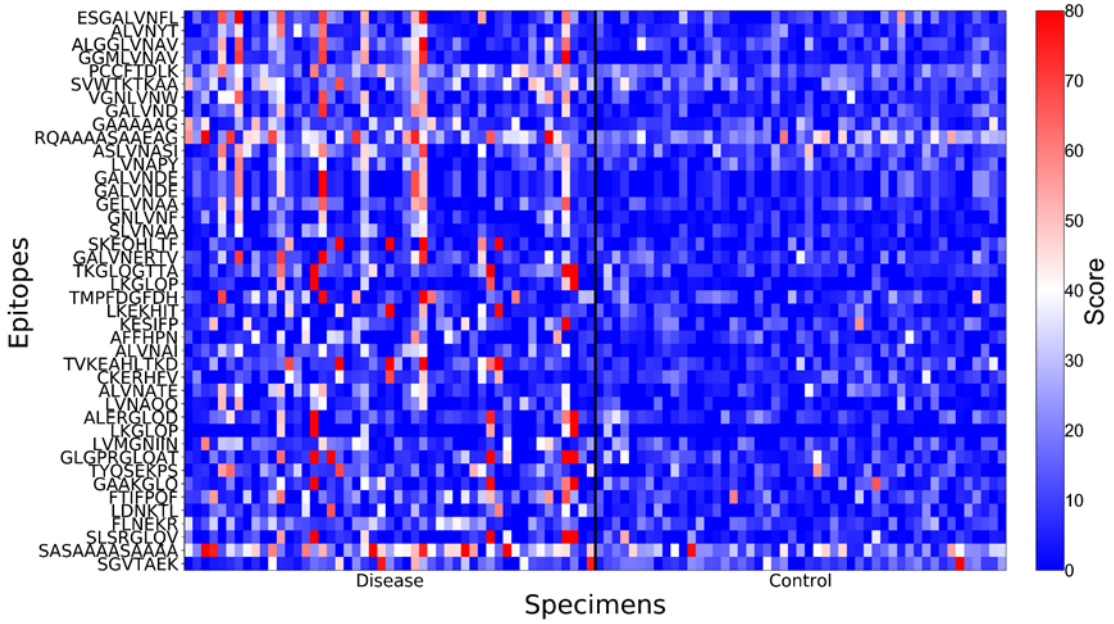


Figure 4.1: Heat map showing epitopes scores for disease and control specimens. The degree to which each specimen bound the 42 AMD-specific epitopes was quantified for all 49 AMD disease and control specimens. These epitopes were highly specific, but not highly sensitive.

We determined binding motifs with MESA for the plausible autoantigens Raftlin, Complement C4-A, and Eyes absent homolog 3 (Figure 4.2). While the epitope logos in Figure 4.2A and Figure 4.2B had clearly defined significant positions, the epitope logo in Figure 4.2C had few significant positions. The regular expression for the binding motif in Figure 4.2C was simply QP, suggesting that this epitope's low prevalence (0.122) may have limited the applicability of substitution analysis.

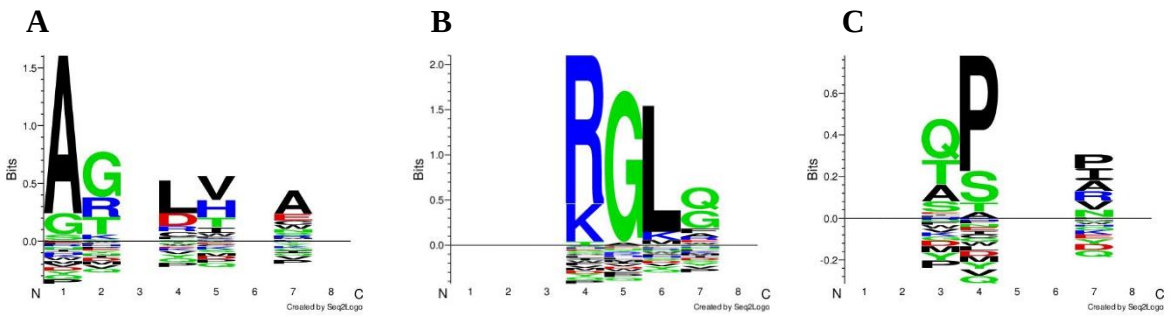


Figure 4.2: Binding motifs for 3 epitopes that had plausible autoantigens. Epitope logos were generated using epitopes (A) GGMLVNAV, (B) ALERGLQD, and (C) TYQSEKPS. The regular expressions for these binding motifs were (A) AGXLV, (B) [RK]GLQ, and (C) QP.

Using the 19 longitudinal specimens, we identified epitopes that were bound at the final timepoint, but were not bound at the initial timepoint. First, we compared the initial and final timepoint for each specimen individually to identify epitopes. Then, we determined consensus epitopes by clustering these 19 sets of epitopes. Through this process, we identified 53 epitopes that were only bound by the final timepoint for at least 3 specimens (top 10 in Table 4.2). However, these epitopes did not appear to correspond to any plausible AMD pathology or match any known autoantigens [159].

Table 4.2: Epitopes that were only bound at the final AMD timepoints. We identified 53 epitopes that were bound by at least 3 final timepoint specimens. Only the 10 epitopes with highest prevalence are shown.

Epitope	Protein	Accession	Prevalence
ITTTTTSTT	CCR4-NOT transcription complex subunit 1	A5YKK6	0.211
RTTTRTTTTTTPTP	Fibronectin type III domain-containing protein 1	Q4ZHG4	0.211
EYIADLYSA	Reticulocalbin-3	Q96D15	0.158
VELATAEAL	Exocyst complex component 3-like protein	Q86VI1	0.158
TTKTPVE	C2 domain-containing protein 5	Q86YS7	0.158
TPFIHNAFK	DNA ligase 4	P49917	0.158
TTSTTSTTI	Mucin-5AC	P98088	0.158
TTTTTTTTTGGI	Cyclin-L1	Q9UK58	0.158
IEADKY	Brefeldin A-inhibited guanine nucleotide-exchange protein 2	Q9Y6D5	0.158
TTSTTAPT	Putative nuclear envelope pore membrane protein POM 121B	A6NF01	0.158

4.2.2 Herpes simplex virus

To identify HSV species-specific epitopes, we analyzed 12 HSV2 specimens and 10 HSV1 specimens. Since these viruses share many of the same proteins in their proteomes [87], HSV1 specimens were appropriate controls for HSV2 specimens and vice-versa. We did not use cross-validation for these specimens since dividing the small specimen sets into training and validation sets would greatly reduce discovery power. To begin, we identified species-specific epitopes in glycoprotein G, which varies significantly between the two species (Figure 4.3) [89]. There was a single HSV1 epitope, PMPSIGLEE, bound by 40% of HSV1 specimens and a single HSV2 epitope, GGPEEFEGAGD, bound by all HSV2 specimens. This HSV2-specific epitope aligned well with previous epitopes found for glycoprotein G2 [90–92] (Table 4.3). Also, this epitope has been validated as an HSV2-specific diagnostic [93,94]. The HSV1-specific epitope was also similar to the previously reported epitope DHTPPMPSIGLE [101]. Interestingly, the two HSV-specific epitopes terminated in an identical 7-mer sequence EGAGDGE (PMPSIGLEEEEEE**EGAGDGE** and GGPEEF**EGAGDGE**) [91]. This suggests that the regions containing these epitopes may be evolutionarily or structurally related targets of the immune system.

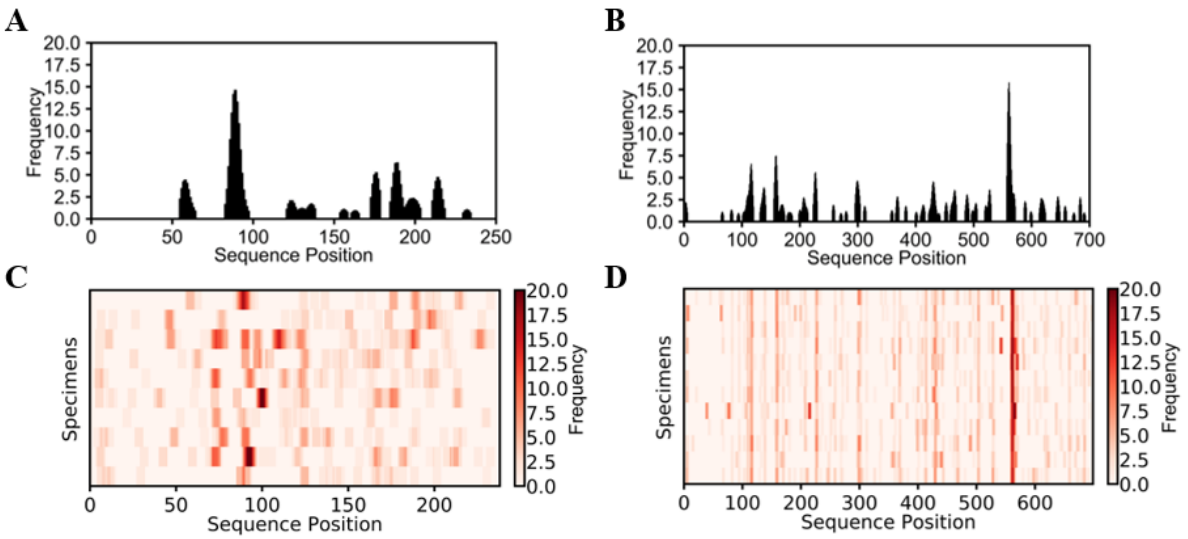


Figure 4.3: K-TOPE identified epitopes for glycoprotein G1 using HSV1 specimens and for glycoprotein G2 using HSV2 specimens. For glycoprotein G1, a representative histogram for a single specimen is shown in (A) and a heat map for all HSV1 specimens is shown in (C). For glycoprotein G2, a representative histogram for a single specimen is shown in (B) and a heat map for all HSV2 specimens is shown in (D). There was a single epitope identified for each protein.

Table 4.3: Alignment of an HSV2-specific glycoprotein G2 epitope with previously reported epitopes.

Peptides	Reference
GG**PEEFEGAGD**	K-TOPE
PEEFEGAGDGEPPEDDDSG	[90]
PPPPEHRGG**PEEFEGAGD**GEPPE	[91]
APPPPEHRGG**PEEFEGAGD**G	[92]

To identify candidate HSV species-specific epitopes, we analyzed the HSV1 and HSV2 proteomes. We identified 30 HSV2 specific epitopes that were 100% specific with prevalence > 30% (Table 4.4). Notably, 11 of these epitopes were present in all HSV2 specimens. K-TOPE identified a glycoprotein C epitope PRTTPTPPQ with 83% prevalence which was contained in a previously identified epitope RNASAPRTTPTPPQPRKATK [101]. Also, a glycoprotein B epitope **KARK**KG**T**SAL shared the sequence KARKXXT with a previously reported epitope **KARK**RK**T**KK [101]. A notable observation was that

five of the 30 HSV2-specific epitopes contained the sequence RXTP. However, two of the antigens that corresponded to these epitopes, envelope glycoproteins H [95] and C [100], were well-described antigens. Since proteins on the outside of a virus, such as the envelope glycoproteins [97], are more plausible targets, we compiled a list of all epitopes in capsid and envelope glycoproteins (Table 4.5). Thus, choosing the most plausible candidate antigens was assisted by additional biological information.

Table 4.4: HSV2-specific epitopes were identified. A total of 30 epitopes were identified that were 100% specific against HSV1.

Epitope	Protein	Accession	Prevalence
GGPEEFEGAGD	Envelope glycoprotein G	P13290	1
PLYARTTPAKF	Tegument protein UL47	P89467	1
VDSQRLTPGGSVS	Tegument protein UL21	P89444	1
KARKKGTSAL	Envelope glycoprotein B	P08666	1
TPLRYACVL	Tegument protein UL47	P89467	1
ANSPWAPVL	mRNA export factor	P28276	1
RYSPLHN	Envelope glycoprotein B	P08666	1
EAMLNDAR	Large tegument protein deneddylase	P89459	1
QRLTPH	Large tegument protein deneddylase	P89459	1
LRYTPAGEV	Envelope glycoprotein H	P89445	1
RTPSMR	Major viral transcription factor ICP4 homolog	P90493	1
LATNNA	Small capsomere-interacting protein	P89458	0.917
LRTNNL	Ribonucleoside-diphosphate reductase small subunit	P69521	0.917
PRTTPTPPQ	Envelope glycoprotein C	Q89730	0.833
HRLYAVVA	Inner tegument protein	P89460	0.833
PSTPAMLNLG	Ribonucleoside-diphosphate reductase large subunit	P89462	0.667
VTKHTALCAR	Large tegument protein deneddylase	P89459	0.583
TRDYAGL	Envelope glycoprotein I	P13291	0.583
RLTVAQ	Envelope glycoprotein I	P13291	0.583
RSLGIA	Protein UL20	P89443	0.583
IRDLARTFA	Thymidine kinase	P89446	0.5
DITAKHRCL	Major capsid protein	P89442	0.5
ETPAQPPRY	Capsid scaffolding protein	P89449	0.5
VSGITPTQ	Tripartite terminase subunit 1	P89451	0.5
HEELYYGPVS	Tegument protein VP22	P89468	0.417
IQDLAYAIV	Ribonucleoside-diphosphate reductase large subunit	P89462	0.417
GPAQRHTY	DNA polymerase catalytic subunit	P89453	0.417
YFEEYAYS	Envelope glycoprotein B	P08666	0.417
LDDFDL	Tegument protein VP16	P68336	0.417
AARLIDALYAEFLGG	Envelope glycoprotein H	P89445	0.333

Table 4.5: HSV2-specific epitopes in plausible antigens. Since the exterior of the virus is more likely to be targeted, we compiled epitopes in glycoproteins and capsid proteins.

Epitope	Protein	Prevalence
GGPEEFEGAGD	Envelope glycoprotein G	1
KARKKGTSAL	Envelope glycoprotein B	1
RYSPLHN	Envelope glycoprotein B	1
LRYTPAGEV	Envelope glycoprotein H	1
PRTTPTPPQ	Envelope glycoprotein C	0.833
TRDYAGL	Envelope glycoprotein I	0.583
RLTVAQ	Envelope glycoprotein I	0.583
DITAKHRCL	Major capsid protein	0.5
ETPAQPPRY	Capsid scaffolding protein	0.5
YFEEYAYS	Envelope glycoprotein B	0.417
AARLIDALYAEFLGG	Envelope glycoprotein H	0.333

In contrast to the numerous HSV2-specific epitopes, only 4 HSV1-specific epitopes were identified and the highest prevalence achieved was only 40% (Table 4.6). One of these epitopes, RIRLPHI, was in the well-described antigen glycoprotein D, and is thus plausibly related to virus neutralization [99]. One possible explanation for the discovery of fewer HSV1-specific epitopes is that the HSV2 specimens had high IgM levels, whereas the HSV1 specimens had high IgG levels. Since high IgM levels occur with severe recurrent herpes infections [164], we would expect the high IgM HSV2 sera to yield more epitopes.

Table 4.6: HSV1-specific epitopes were identified. Only 4 epitopes were identified that were 100% specific against HSV2.

Epitope	Protein	Accession	Prevalence
RIRLPHI	Envelope glycoprotein D	Q69091	0.4
PMPSIGLEE	Envelope glycoprotein G	P06484	0.4
CAAFVNDYSLV	Major capsid protein	P06491	0.3
EMADTFLDT	ICP47 protein	P03170	0.3

We sought to determine whether the HSV2-specific epitopes were contained in proteins that differed between the HSV species [89]. We determined 8 HSV2-specific epitopes with sequences that were contained in both HSV proteomes (Table 4.7). Our

analysis suggested that these epitopes were only targeted by HSV2 specimens, despite their presence in the HSV1 proteome. Thus, even regions that are conserved between species could serve as species-specific targets.

Table 4.7: Eight HSV2-specific epitopes were also in the HSV1 proteome.

Epitope	Protein	Accession	Prevalence
PLYARTTPAKF	Tegument protein UL47	P89467	1
TPLRYACVL	Tegument protein UL47	P89467	1
ANSPWAPVL	mRNA export factor	P28276	1
QRLTPH	Large tegument protein deneddylase	P89459	1
LRTNNL	Ribonucleoside-diphosphate reductase small subunit	P69521	0.917
PSTPAMLNLG	Ribonucleoside-diphosphate reductase large subunit	P89462	0.667
YFEEYAYS	Envelope glycoprotein B	P08666	0.417
LDDFDL	Tegument protein VP16	P68336	0.417

Binding motifs were determined for three HSV1-specific epitopes and three HSV2-specific epitopes (Figure 4.4). The regular expression for the PLYARTTPAKF binding motif was [YM]X**RX**[**T**DL]**P** (Figure 4.4B). This regular expression contained RXTP, which was noted as being present in several HSV2-specific epitopes (Table 4.4). In the epitope logo for KARKKGTSAL (Figure 4.4C), the last 6 positions were important to binding. However, the K-TOPE epitope **KARK**KGTSAL overlapped with the Risinger et al. epitope (**KARK**RKTKK) at the first 4 positions [101]. Thus, the epitope logo revealed that the K-TOPE epitope was not plausibly related to the Risinger et al. epitope. For the HSV1-specific binding motifs, approximately 30-45% of the positions in the epitopes were deemed insignificant for binding. This result implies that these insignificant positions could potentially be mutated without a concomitant decrease in binding.

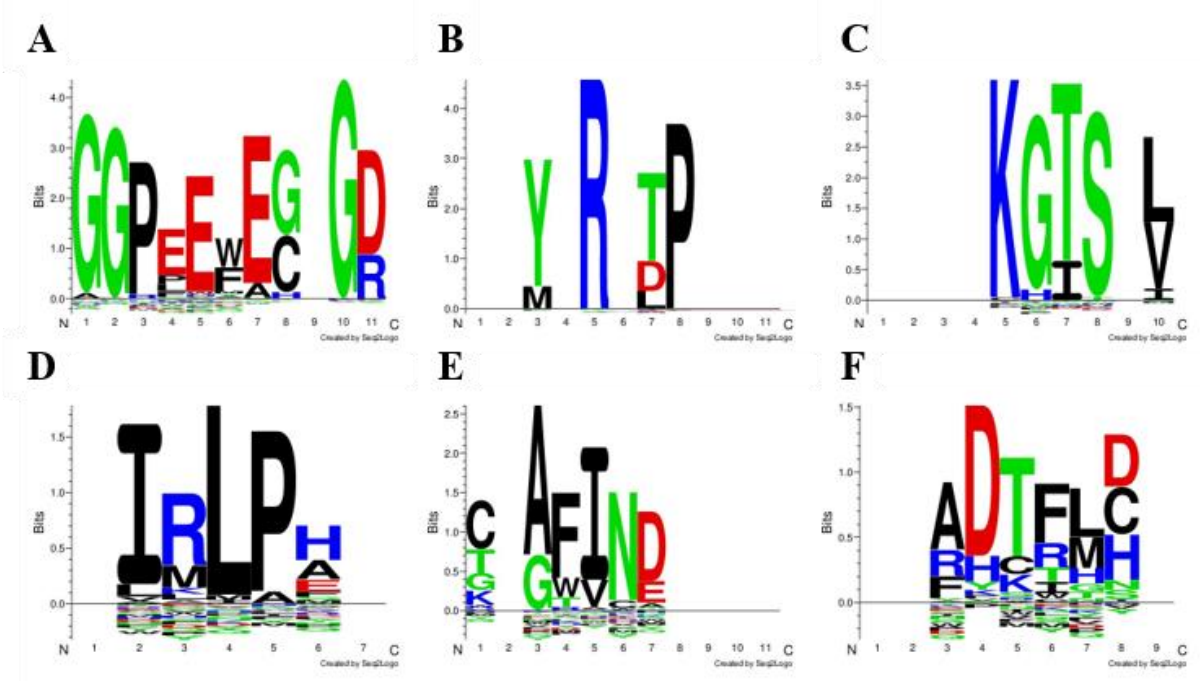


Figure 4.4: Binding motifs were determined for 3 HSV1-specific epitopes and 3 HSV2-specific epitopes. Epitope logos (A), (B), and (C) correspond to HSV2-specific epitopes. Epitope logos (D), (E), and (F) correspond to HSV1-specific epitopes. MESA was used with the epitopes (A) GGPEEFEGAGD, (B) PLYARTTPAKF, (C) KARKKGTSAL, (D) RIRLPHI, (E) CAAFVNDYSLV, and (F) EMADTFLDT. The regular expressions for these binding motifs were (A) GGP[EP]E[WF]E[GC]XG[DR], (B) [YM]XRX[TDL]P, (C) KG[TI]SX[LV], (D) I[RM]LPH, (E) CX[AG][FW][IV]ND, and (F) ADTF[LM][DHC].

4.2.3 Chagas disease

We identified epitopes that were present in at least 10% of 45 disease specimens and no more than 5% of 30 control specimens. Then, we re-calculated the prevalence and specificity of these epitopes using a validation set of 45 disease specimens and 30 control specimens. Epitopes which didn't meet the 10% prevalence and 95% specificity thresholds in the validation set were removed.

We first sought to characterize the diagnostically important Chagas antigen, trypomastigote small surface antigen (TSSA) [110,111,113,165]. In analyzing TSSA, we

identified the epitope ENKPATGEA, which had a prevalence of 0.178 and a specificity of 0.967 in the validation set. This epitope had significant overlap with previously identified epitopes (Table 4.8). Taken together, these three epitopes implied that the core TSSA epitope was approximately KPATGE. This was confirmed by the epitope logo (Figure 4.5), which has the regular expression [PW]XTGE.

Table 4.8: Alignment of a TSSA epitope with previously reported epitopes.

Peptides	Reference
E N **K P A T G E** A	K-TOPE
T S S T P P S G T E N **K P A T G E**	[24]
T S S T P P S G T E N **K P A T G E** A P S Q	[110]
K P A T G E A P S Q	[111]

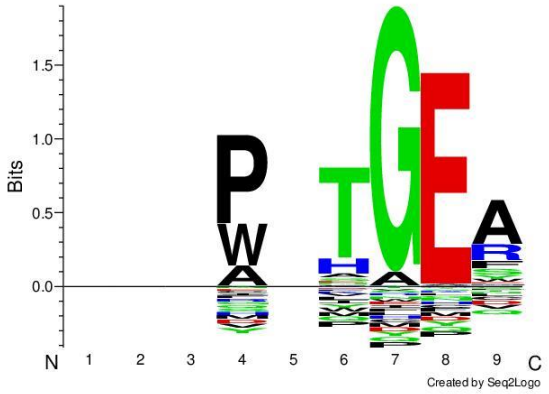


Figure 4.5: Binding motif determined for the TSSA epitope ENKPATGEA. The binding motif had a regular expression of [PW]XTGE, which is similar to previously reported epitopes.

To identify epitopes and antigens related to the progression of Chagas disease, we used K-TOPE with the *T. cruzi* proteome (10,770 proteins). Through this analysis, we identified 222 Chagas-specific *T. cruzi* epitopes (top 25 in Table 4.9). Notably, there were 11 epitopes which were targeted by greater than half of the disease specimens. Many of the proteins that were identified are known antigens such as Mucin TcMUCII [114,116], trans-

sialidase [117], dispersed gene family protein 1 (DGF-1) [114,118], and Mucin-associated surface protein (MASP) [120–122].

Table 4.9: The 25 Chagas-specific epitopes with the highest prevalence. A total of 222 epitopes were identified that were Chagas-specific. *Previously validated antigens

Epitope	Protein	Accession	Prevalence	Specificity
RAGDKVE	Uncharacterized protein	Q4DQL9	0.8	0.967
EAGDKVQG	Trans-sialidase, putative*	Q4DVK2	0.733	0.967
AERMRAIE	Uncharacterized protein	Q4DEQ5	0.644	1
SRLREIDG	Putative mucin EMUCt-4*	Q962H7	0.622	1
SRFREIDG	Mucin TcMUCII, putative*	Q4D1D0	0.622	1
SRLREIDGS	Mucin TcMUCI, putative*	Q4DR54	0.6	1
APAAGGFGSA	Uncharacterized protein	Q4D768	0.6	0.967
RLREIDGS	Mucin TcMUCII, putative (Fragment)*	Q4D5F8	0.578	1
PSRLREIDG	Mucin-like protein*	O61045	0.556	1
EFRQIDT	Uncharacterized protein	V5B8U7	0.533	1
DEGFGWVER	40S ribosomal protein SA	Q4CQ63	0.511	1
LAGGFGEL	Trans-sialidase, putative*	Q4CWN6	0.489	1
CAGDKN	Uncharacterized protein	Q4DLH9	0.467	1
MRLIDAVAR	Uncharacterized protein	Q4D0V6	0.467	1
VAGDKC	Uncharacterized protein	Q4DT98	0.467	0.967
IRAFRLIDV	Centrin, putative	Q4DBB8	0.444	1
EAGGFGVL	Glucokinase 1, putative	Q4E4E1	0.444	1
SGGFGR	Mucin-associated surface protein (MASP), putative*	Q4CZA8	0.444	1
VGGFGTG	Succinyl-CoA:3-ketoacid-coenzyme A transferase	Q4D0L3	0.444	1
KAGGFGNRVV	Uncharacterized protein	Q4DIM5	0.422	1
MRQIDEL	Prostaglandin F synthase	Q4DJ07	0.422	0.967
RGGFGASA	Uncharacterized protein	Q4DAX7	0.422	0.967
KIRAIEA	Ribose 5-phosphate isomerase	Q4CQE2	0.422	0.967
FEGGFGS	Dispersed gene family protein 1 (DGF-1), putative*	Q4DTA7	0.4	1
YGGFGAS	Acyl-CoA dehydrogenase, putative	Q4DLR8	0.4	1

Of the 39 Chagas-specific motifs identified by the antibody repertoire profiling company SerImmune [115], nine were similar to K-TOPE epitopes (Table 4.10). Additionally, using K-TOPE, three SerImmune motifs that could not be associated with antigens were associated with surface antigen 2 (B13) [119], Mucin TcMUCII, and DGF-1. Notably, 81% of K-TOPE epitopes could not be matched to SerImmune motifs, suggesting that these epitopes may be novel.

Table 4.10: Comparison between SerImmune motifs and K-TOPE epitopes. Of the 39 SerImmune motifs, 9 were similar to K-TOPE epitopes. The SerImmune motifs and their suspected antigens are shown in bold. For some motifs, the list of matching K-TOPE epitopes was truncated. Three of the SerImmune motifs were newly associated with putative antigens.

SerImmune Motif	Suspected Antigens	K-TOPE epitope	K-TOPE antigen
[RK]MRXID	**None**	AERMRAIE	Uncharacterized protein
		SRLREIDG	Putative mucin EMUCt-4
		SRFREIDG	Mucin TcMUCII, putative
		SRLREIDGS	Mucin TcMUCI, putative
ETXIPXE	**Complement regulatory protein, Trans-sialidase, FL-160-1 epitope, 05M3-like kinesin**	PGETKIPSE	Trans-sialidase, putative
		GETKIPSES	Flagellum-Associated Protein (Fragment)
RXSPYX[IL]F	**Kinetoplast DNA-associated protein 3**	SPYSIFLQE	Kinetoplast DNA-associated protein 3
PQXQH[ED]	**Helicase, putative, Phosphatidylinositol 3-kinase**	PEIQHD	Uncharacterized protein
PXXGGFG	**None**	APAAGGFGSA	Uncharacterized protein
		SGGFGR	Mucin-associated surface protein (MASP), putative
		GGFGQE	ATP-dependent RNA helicase, putative
		DGGFGG	Uncharacterized protein
HYEWA	**Lanosterol cyclase, Terpene cyclase/mutase family member**	EHYEWAAG	Terpene cyclase/mutase family member
GREXDG	**Mucin-associated surface protein (MASP), Trypanothione synthetase-like protein**	GREIDD	Trans-sialidase, putative
		DGREID	Glutaminyl cyclase, putative
A[KR]AG[DN]K	**None**	RAGDKVE	Uncharacterized protein
		CAGDKN	Uncharacterized protein
		KAGDKKR	Uncharacterized protein
		VAGNKQ	Trans-sialidase, putative
F[RN]XIN[RQ]	**Dynein heavy chain, Eukaryotic translation initiation factor 3 subunit 8**	EFRQIDT	Uncharacterized protein
		LREIDRI	Uncharacterized protein
		RREIDR	Paraflagellar rod protein, putative
		FRQIEI	Aminoalcohol phosphotransferase, putative

To provide additional evidence that K-TOPE epitopes were Chagas-specific, we compared K-TOPE epitopes to an ELISA diagnostic for Chagas disease [116]. Four members of the seven-peptide diagnostic panel matched K-TOPE epitopes (Table 4.11). Thus, the Chagas-specific epitopes identified by K-TOPE were well-corroborated by previous research.

Table 4.11: K-TOPE epitopes matched a diagnostic panel of peptides. Four out of seven peptides using in a diagnostic ELISA for Chagas disease matched K-TOPE epitopes. The ELISA peptides and their antigens are shown in bold.

ELISA Peptide	ELISA Antigen	K-TOPE epitope	K-TOPE antigen
APFGQAAAGDKPSPF	**b13 / Ag2 / CA-2 / PEP2**	PSPFGQAAAG	Surface antigen 2 (CA-2), putative
EPKSAEPKPAEPKSA	**TcD / Ag13**	AEPKSAEPK	Trans-sialidase, putative
		STPAEPKPA	Trans-sialidase, putative
TTNAPSRLREIDGSL	**Mucin TcMUCII**	SRLREIDG	Putative mucin EMUCt-4
		SRFREIDG	Mucin TcMUCII, putative
		SRLREIDGS	Mucin TcMUCI, putative
		RLREIDGS	Mucin TcMUCII, putative (Fragment)
		PSRLREIDG	Mucin-like protein
DSAKGKATGSSAGED	**Trans-sialidase**	AKGKATGS	Trans-sialidase, putative

Binding motifs were determined for three prevalent Chagas-specific epitopes (Figure 4.6). The epitope logos showed that for these epitopes, only 4 or 5 amino acids were significant, such as GGFG for the epitope APAAGGFGSA.

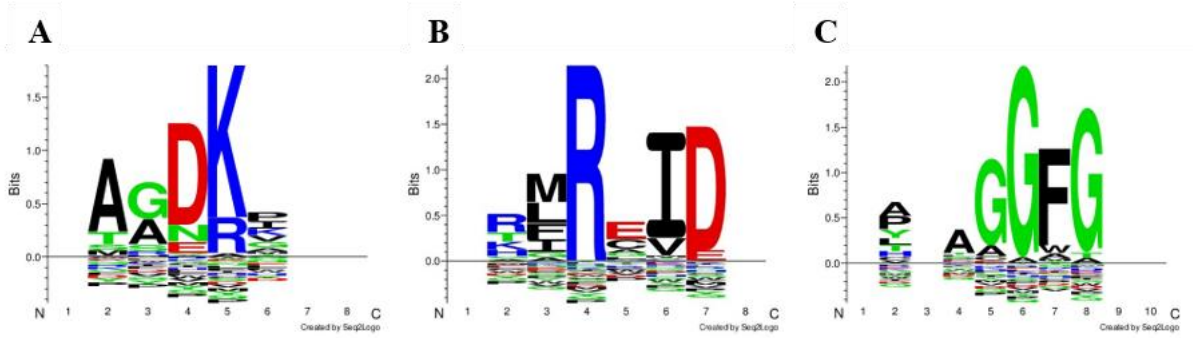


Figure 4.6: Binding motifs determined for three Chagas-specific epitopes. Epitope logos were generated using epitopes (A) EAGDKVQG, (B) RLREIDG, and (C) APAAGGFGSA. The regular expressions for these binding motifs were (A) AGD[KR], (B) MRXID, and (C) GGFG.

To assess the limitations of using separate training and validation specimen sets, we performed a 2-fold cross validation scheme. In this scheme, we identified Chagas-specific epitopes using a training and validation set, and then switched the training and validation sets to identify additional epitopes. Finally, we combined the two sets of epitopes and removed any redundancy. Through this analysis, we determined that there was a 2.7% difference in the number of epitopes identified using either training set. After combining the two sets, there was a 17.1% increase in the number of epitopes (from 222 to 260) as compared to using separate training and validation sets. Thus, regardless of which half of the data was used for training, most of the epitopes identified were the same. These results suggest that using separate training and validation specimen sets, rather than cross-validation, appears to sufficiently characterize this dataset.

Since there is evidence of autoimmunity in Chagas disease [104], we analyzed the human proteome (19,603 proteins) for epitopes. This analysis yielded 1,084 epitopes (top 25 in Table 4.12). By comparing these epitopes and their antigens to known autoantigens, we showed that two of the K-TOPE antigens, mi-2b and MDA-5, were implicated in myositis [159]. Since there is cardiac involvement in myositis [166], these two autoantigens could plausibly exacerbate the cardiomyopathy exhibited in Chagas disease. Although, it is important to note that all candidate autoantigens could be investigated since sometimes relevant autoantigens show no obvious relation to pathological phenotypes, such as the autoantigen topoisomerase in systemic sclerosis [167].

Table 4.12: Candidate autoantigens for Chagas disease. Analyzing the human proteome revealed 1,084 Chagas-specific epitopes. Only the 25 epitopes with highest prevalence are shown.

Epitope	Protein	Accession	Prevalence	Specificity
SKMRAIDLQ	Nesprin-2	Q8WXH0	0.733	0.967
RAGDKN	Protein FAM71B	Q8TC56	0.644	1
ESAQAGDKC	DNA replication licensing factor MCM6	Q14566	0.644	1
GFGAAGGFGGR	Keratin, type II cytoskeletal 2 epidermal	P35908	0.6	1
VISAFAGDKD	Cerebellin-4	Q9NTU7	0.578	1
PGGFGS	Replication protein A 32 kDa subunit	P15927	0.578	0.967
FPGGFGAA	ES1 protein homolog, mitochondrial	P30042	0.578	0.967
SGGFGS	Deoxyuridine 5'-triphosphate nucleotidohydrolase, mitochondrial	P33316	0.556	1
GGGFGGFGS	Keratin, type II cytoskeletal 1	P04264	0.556	0.967
SGGFGS	Serine/threonine-protein kinase pim-1	P11309	0.533	1
AAGGFGRA	Putative solute carrier family 22 member 31	A6NKX4	0.511	1
SGGFGSR	Keratin, type II cytoskeletal 4	P19013	0.511	1
FRQIDSG	Ethanolamine-phosphate cytidylyltransferase	Q99447	0.511	1
APAGGFGGFGT	Nucleoporin p54	Q7Z3B4	0.511	1
VKMRGIDF	Fructose-2,6-bisphosphatase TIGAR	Q9NQ88	0.511	1
SAKIRAIE	Rabenosyn-5	Q9H1K0	0.511	0.967
VPGGFGVR	CTP synthase 1	P17812	0.511	0.967
VPQAGGFGC	Plasma membrane calcium-transporting ATPase 3	Q16720	0.489	1
SSASGGFGS	Nuclear pore complex protein Nup214	P35658	0.489	1
LSGGFGSR	Keratin, type II cytoskeletal 71	Q3SY84	0.489	1
NRLREIDE	Fibrous sheath-interacting protein 1	Q8NA03	0.489	1
SGGFGAF	Zinc transporter ZIP11	Q8N1S5	0.489	1
RGGFGS	40S ribosomal protein S2	P15880	0.489	1
AGGFGGGGFGG	Keratin, type II cytoskeletal 1b	Q7Z794	0.489	0.967
LAGGFGF	Adenylate cyclase type 5	O95622	0.467	1

Patients with Chagas infections often have coinfections with *Leishmania major (L. major)* [123], thus there is a need for a Chagas diagnostic which is specific against *L. major*. Approximately 1000 [168] out of 8300 [169] *L. major* genes are not found in *T. cruzi*. Due to this considerable genetic similarity (88%), it was important to remove epitopes bound by leishmaniasis specimens from the list of Chagas-specific epitopes. To remove Chagas-specific epitopes that may also be associated with *L. major* antigens, we used the *L. major* proteome (7,863 proteins) to identify 929 epitopes bound by at least 2 of 12 *L. major* specimens and none of 12 Chagas control specimens. Then, we removed any epitopes from

the 222 Chagas-specific epitopes (Table 4.9) that matched the 929 *L. major* epitopes. Through this analysis, we identified 197 epitopes (top 10 in Table 4.13), a reduction of approximately 10% from the 222 epitopes identified earlier (Table 4.9). Notably, the prevalent epitopes RAGDKVE and APAAGGFGSA (Table 4.9), were removed by this analysis. This suggests that seemingly disease-specific epitopes should be validated against similar diseases.

Table 4.13: Chagas-specific epitopes that do not cross-react with *L. major*. Epitopes RAGDKVE and APAAGGFGSA from Table 4.9 were removed in this analysis. Only the 10 epitopes with highest prevalence are shown.

Epitope	Protein	Accession	Prevalence	Specificity
EAGDKVQG	Trans-sialidase, putative	Q4DVK2	0.733	0.967
AERMRAIE	Uncharacterized protein	Q4DEQ5	0.644	1
SRLREIDG	Putative mucin EMUCt-4	Q962H7	0.622	1
SRFREIDG	Mucin TcMUCII, putative	Q4D1D0	0.622	1
SRLREIDGS	Mucin TcMUCI, putative	Q4DR54	0.6	1
RLREIDGS	Mucin TcMUCII, putative (Fragment)	Q4D5F8	0.578	1
PSRLREIDG	Mucin-like protein	O61045	0.556	1
EFRQIDT	Uncharacterized protein	V5B8U7	0.533	1
DEGFGWVER	40S ribosomal protein SA	Q4CQ63	0.511	1
LAGGFGEL	Trans-sialidase, putative	Q4CWN6	0.489	1

4.3 Discussion

We have demonstrated that K-TOPE and MESA can be used to identify and characterize disease-specific epitopes and antigens. We identified novel and previously reported autoantigens for an autoimmune disease (AMD), a viral disease (HSV), and a parasitic disease (Chagas disease). Since we could apply these approaches to a broad spectrum of diseases, we infer that these methods could be applied to numerous additional diseases. Importantly, even researchers without antibody repertoire analysis expertise could integrate K-TOPE and MESA into their analytical pipelines to discover disease-specific

epitopes and antigens. Thus, these algorithms could greatly enhance the rate at which disease-specific epitopes and antigens are discovered.

Analyzing three different diseases revealed multiple features of this approach that could help improve future epitope discovery. The selection of specimens, specifically sample size, is extremely important to study design. For AMD, it is likely that we only identified low prevalence epitopes because we used specimens in various states of AMD and used heterogenous controls. Unfortunately, using a small sample size for the HSV analysis reduces the probability that our results will generalize to a larger population. One computational technique that can counter sample size limitations is cross-validation [170]. In the case of AMD, cross-validation revealed a significant number of additional epitopes. Although the effect of cross-validation was less pronounced for Chagas disease which had a larger sample size.

An important aspect of these approaches is that epitopes and antigens need to be validated experimentally or using previous studies. Experimental validation could consist of spotting antigens on to a microarray or ELISA to determine if they can differentiate between disease and control specimens [171]. For HSV and Chagas disease, we validated epitopes using previous studies. However, it was difficult to know *a priori* whether a non-validated epitope was novel or spurious. In general, since studies use different specimens, experiments, and computational analyses, it is unlikely for the results of two studies to completely coincide. An example of these differences was shown in the comparison of motifs identified by SerImmune to epitopes identified by K-TOPE. SerImmune used computational techniques that could identify motifs corresponding to non-linear and non-protein epitopes [43]. Therefore, we would not expect SerImmune motifs to completely agree with K-TOPE

epitopes, which are exclusively derived from protein sequences. Using MESA also aids in validating epitopes since it can help determine if two epitopes are equivalent. In the case of KARKKGTSAL (Figure 4.4C), the regular expression of the epitope suggested that it corresponded to a different antibody specificity than the previously reported epitope, KARKRKTKK [101]. In discovering epitopes, it is important to realize that the minimum prevalence chosen will largely determine how many epitopes are discovered. For instance, although there were 929 *L. major*-specific epitopes and 30 HSV2-specific epitopes, both groups had 12 specimens. This difference can be explained by the choice of a minimum prevalence of 2 specimens for *L. major* and 4 specimens for HSV2.

Biological understanding of a disease can also aid epitope and antigen discovery. Since HSV1 and HSV2 are so closely related [87], analyzing both species was critical to establishing whether an epitope was truly specific. Unexpectedly, we demonstrated that even epitopes present in the conserved regions of both species' proteomes could be species-specific. This is likely due to differences in the structure and post-translational modifications of the proteins. Knowing that *T. cruzi* and *L. major* often cause coinfections [123], we were able to identify Chagas-specific epitopes that were not bound by leishmaniasis specimens. This example shows that to develop a clinically useful diagnostic, it is often necessary to incorporate epidemiological information into a study.

The approaches used here have advantages over similar epitope identification schemes. While proteome-derived peptide libraries have been used to identify disease-specific epitopes [21,24], these methods lack the flexibility to examine multiple proteomes. For instance, separate libraries would be required to analyze both HSV1 and HSV2. Also, for Chagas disease, it would be impractical to construct a non-random library that covers the entire *T.*

cruzi proteome. Using K-TOPE, we were able to identify epitopes towards the parasites *T. cruzi* and *L. major* using the same experimental procedure. While random libraries have also been used to identify disease-specific epitopes [34], it is often difficult to connect these epitopes to antigens in a statistically rigorous manner. By integrating epitope and antigen identification, K-TOPE automatically associates every epitope with an antigen. Finally, to identify important binding positions in an epitope, most approaches construct a separate library for each epitope [28,39,44]. With MESA, multiple epitopes are analyzed *in silico* to identify the most important binding positions and amino acid preferences.

In this study, we identified epitopes and antigens that could be used for a variety of medical applications. A collection of disease-specific epitopes could be combined into a multi-epitope peptide for vaccination [101]. Or, novel antigens could be used in a conventional ELISA to diagnose diseases with high sensitivity [172]. Also, an epitope that is known to contribute to a disease's pathology could be targeted by a therapeutic monoclonal antibody [145]. By applying this approach to any of the numerous diseases that involve antibody responses, it could be possible to more effectively prevent, diagnose, and treat diseases.

4.4 Materials and methods

4.4.1 Strains and reagents

All library screening experiments used *E. coli* strain MC1061 with display vector pB33eCPX. All sera (n=301) were obtained as deidentified specimens from biobanks according to institutional guidelines, (Biosafety authorization numbers #201417, #201713), and handled according to CDC-recommended BSL2 guidelines. We obtained 12 HSV2 serum specimens from BioreclamationIVT and 10 HSV1 serum specimens from Discovery

Life Sciences. The 117 AMD specimens were collected as part of the AREDS study [158] and included 49 disease specimens, 49 control specimens, and 19 final timepoint disease specimens. The 90 Chagas disease specimens, 60 control specimens, and 12 leishmaniasis specimens were collected by the CDC and screened by the antibody repertoire profiling company SerImmune.

4.4.2 Screening and sequencing bacterial peptide display libraries

Random bacterial peptide display libraries were screened as previously described [124]. Briefly, we combined a 12-mer peptide display library with 1:100 diluted serum and used magnetic Protein A/G beads (Thermo Scientific Pierce) to sort for library members with antibody-binding peptides. We prepared amplicons from the sorted library members and sequenced the amplicons using next generation sequencing (NextSeq). The algorithms used to identify peptide sequences and evaluate k-mers are outlined in Chapter 2.

4.4.3 Protein database searches

We identified protein sequences using UniProt or with the Biopython module [142]. Protein accessions are noted in the presented tables. The accessions for glycoprotein G1, glycoprotein G2, and TSSA were P06484, P13290, and D0VAV8. The random proteins used to assign statistical significance with K-TOPE were obtained through a UniProt search of "reviewed:yes". Human proteome searches used a UniRef search of "uniprot:(fragment:no reviewed:yes organism:"Homo sapiens (Human) [9606]" proteome:up000005640) AND identity:0.9" which yielded 19,603 proteins. HSV analysis used a UniProt search of "reviewed:yes AND organism:"Human herpesvirus 1 (strain 17) (HHV-1) (Human herpes simplex virus 1) [10299]" AND proteome:up000009294" for HSV1, yielding 73 proteins and a Uniprot search of "reviewed:yes AND organism:"Human herpesvirus 2 (strain HG52)

(HHV-2) (Human herpes simplex virus 2) [10315]" AND proteome:up000001874" for HSV2, yielding 72 proteins. The *T. cruzi* proteome search used a UniRef search of "uniprot:(fragment:no organism:"Trypanosoma cruzi (strain CL Brener) [353153]" proteome:up000002296) AND identity:0.9" and yielded 10,770 proteins. The *L. Major* proteome search used a UniRef search of "uniprot:(fragment:no organism:"Leishmania major [5664]" proteome:up000000542) AND identity:0.9" and yielded 7,863 proteins.

4.4.4 Epitope identification

Epitopes were identified using K-TOPE (described in Chapter 2). All analyses used "disease" group specimens to identify epitopes and "control" group specimens to subtract epitopes. Epitopes were identified in the disease group that met the epitope percentile cutoff (95%) and the minimum prevalence (varied by disease). Then all disease epitopes were evaluated in the control group. For an epitope to be considered disease-specific, its score had to be below the epitope percentile cutoff (80%) in a proportion of the control specimens, as determined by the specificity cutoff (varied by disease). These disease-specific epitopes were then evaluated in a validation set to recalculate their prevalence and specificity. Epitopes that did not meet the sensitivity and specificity thresholds in the validation set were removed from the disease-specific epitope list. For the Chagas analysis, the PAM30 similarity matrix was used to remove redundant epitopes with a cutoff of 10 (see Chapter 2.4.6).

For the 2-fold cross validation analysis, disease-specific epitopes were identified and validated as stated above, except that the analysis was repeated by switching the training and validation sets. Then, the two disease-specific epitope lists were combined and redundancy was removed using the PAM30 matrix. Finally, the prevalence and specificity of the epitopes in the combined list were re-evaluated in all disease and control specimens. To identify

"disease progression" epitopes for AMD, epitopes were identified separately using each pair of initial and final timepoint specimens. Therefore, all epitopes that were at both time points for a specimen were removed. Then, all disease progression epitopes were clustered to determine consensus epitopes. Finally, we removed any consensus epitopes that were bound at the initial timepoints, which may have emerged due to the clustering process.

To identify HSV2-specific epitopes that were also in the HSV1 proteome, we identified epitopes that exactly matched a subsequence in an HSV1 protein. We determined the similarity between previously reported Chagas-specific epitopes and K-TOPE epitopes using the PAM30 matrix. To identify Chagas-specific epitopes that were not also bound by *L. major* specimens, we removed any *L. major*-specific epitopes from the list of Chagas-specific epitopes.

We compared K-TOPE autoantigens to previously described autoantigens by identifying 76 UniProt proteins that matched 52 previously described autoantigens [159]. We then compared the names of the autoantigen proteins to the names of K-TOPE proteins to identify matches.

4.4.5 Epitope logo generation

Epitope logos were generated using MESA (Chapter 2). All analyses used 6-mers, a frequency cutoff of 0.15, and a minimum enrichment threshold of 20%. AMD, HSV, and Chagas analyses used score thresholds of 10%, 10%, and 5% respectively. Chagas used a lower score threshold since its analysis used a higher number of specimens.

4.4.6 Data visualization

All figures were created using the Matplotlib python module [144] or Seq2Logo [155].

5 Conclusions

5.1 Summary

5.1.1 Identification of linear protein epitopes

For many medical applications, it is important to identify epitopes in target antigens. Approaches that seek to identify the epitopes targeted by antibody repertoires often use peptide libraries (Chapter 1). However, these methods generally cannot explicitly link linear epitopes to their corresponding protein antigens. We developed the K-mer Tiling of Protein Epitopes (K-TOPE) algorithm (Chapter 2) to address the insufficiency of existing approaches. The main input to this algorithm is antibody-binding peptides, which are determined by screening a surface-displayed random peptide library with serum. Using a large random library provides a rich dataset, allowing for the analysis of many antibody responses. One of the defining features of K-TOPE is that it tiles antigen sequences of interest into overlapping k-mers. The k-mers can then be evaluated for enrichment in the set of antibody-binding peptides and compiled to reveal epitopes. By scaling this approach up to whole proteomes and large specimen sets, K-TOPE can characterize how a population binds to whole organisms.

We first determined that this approach could identify the signature of a single antibody among the numerous antibodies in serum by spiking monoclonal and polyclonal antibodies into serum. In this case, the epitopes that K-TOPE identified clearly matched those of previous studies. To identify commonly bound epitopes, we applied this method to the causative agent of the common cold, *Rhinovirus A*. Through this effort, we identified three epitopes that were bound by 83% of a set of 250 serum specimens. We then extended

this analysis to all viral proteomes, consisting of nearly 3,000 proteins from 400 viral taxa, to identify seven enterovirus epitopes and 5 Epstein-Barr virus epitopes recognized by >30% of 250 specimens. Interrogating the common bacterial genera of *Staphylococcus* and *Streptococcus* revealed six epitopes bound by >40% of the specimens. Finally, these prevalent bacterial and viral epitopes matched previously described epitopes with statistical significance.

We validated K-TOPE using antigens from ubiquitous pathogens that should be bound by any human population, rather than focusing on antibody responses related to a specific active infection. Therefore, repeating this analysis with a different set of specimens should generate similar "public epitopes". Since many of the epitopes identified by K-TOPE have not been previously described, it is likely that K-TOPE identified novel epitopes. Unfortunately, validation using previous studies can be difficult since epitope databases often lack quantitative information for epitopes. An important limitation of K-TOPE is that it often identifies multiple antigens that contain similar epitopes. The identification of multiple candidate antigens is a consequence of extensive protein sequence similarity in nature. Therefore, differentiating between candidate antigens requires assaying whole proteins.

5.1.2 Characterization of epitope binding motifs

Generally, approximately five amino acids dominate binding in an epitope [46]. However, the epitopes generated by K-TOPE are always longer than five amino acids. Therefore, it was important to develop a tool that could identify which positions in an epitope were crucial to binding. To address this need, we developed Multiplexed Epitope Substitution Analysis (MESA) (Chapter 3). In addition to identifying important binding positions, MESA determines which amino acids are preferred at each position. This approach

was based on exhaustive substitution studies using targeted microarrays [44]. However, in contrast to microarray-based approaches, MESA utilizes random peptide library screening to enable exhaustive substitution on multiple epitopes in parallel. MESA is similar to K-TOPE in that it divides a sequence into overlapping k-mers. In contrast to K-TOPE, MESA then substitutes each position in the k-mers with all amino acids. By using the relative enrichments of the k-mers and their substitutions, MESA identifies binding motifs which indicate important binding positions and amino acid preferences.

To validate MESA, we generated binding motifs for epitopes and compared them to motifs generated by an alternative computational method, MEME. We made this comparison since both MEME and MESA seek to identify binding motifs, although using completely different approaches. We analyzed a serum specimen with spiked-in monoclonal antibodies and identified binding motifs that were highly similar to MEME motifs. Through analyzing multiple serum specimens, we characterized binding in common pathogen epitopes. One advantage of MESA is that it can use full sets of antibody-binding peptides, whereas MEME can only use <1% of all peptides. Due to this difference, we identified a previously described binding motif which could not be identified by MEME.

One of the most vital advantages of MESA is that it can determine multiple binding motifs in parallel. To our knowledge, all existing schemes for determining binding motifs require starting with a non-random library. In contrast, MESA uses a random library to enable the determination of binding motifs for epitopes of viral, bacterial, autoimmune, or unknown origin. An additional advantage of using a random library is that the same set of antibody-binding peptides can be used to identify epitopes with K-TOPE and generate binding motifs with MESA. It is important to determine binding motifs for K-TOPE epitopes

for the cases where an antibody optimally binds to a sequence that differs from the antigen sequence. In these situations, MESA can identify the optimal amino acids for binding at each position in the epitope. MESA should only be used when a significant percentage of specimens binds an epitope. If a group of specimens does not have homogenous binding to an epitope, MESA may inaccurately characterize the binding motif. MESA also requires large datasets of antibody-binding peptides with a high percentage of all possible 5- or 6-mers represented.

5.1.3 Identifying and characterizing disease-specific epitopes and antigens

To further validate K-TOPE and MESA, we analyzed diseases that feature significant antibody responses (Chapter 4). The diseases we analyzed were age-related macular degeneration (AMD), herpes simplex virus (HSV), and Chagas disease. By analyzing autoimmune, viral, and parasitic diseases, we demonstrated that these approaches could be applied to diseases with a wide range of etiologies. The autoimmune disease AMD is the leading cause of blindness in the developed world and is accepted to involve immune dysregulation and the generation of autoantibodies. Viral infection by HSV involves two closely related species, HSV1 and HSV2, and causes cold sores and genital ulcers. Finally, Chagas disease is a parasitic infection by *T. cruzi* that can lead to cardiomyopathy and digestive megasyndromes, and is exacerbated by autoimmunity. This parasite often has coinfections with *L. major*, the causative agent of leishmaniasis. We sought novel antigens for these diseases that could be used in medical applications.

For all three diseases, we identified previously reported and novel epitopes and antigens. We identified 42 AMD-specific epitopes in candidate autoantigens, three of which had a plausible connection to the progression of disease. Also, using AMD specimens that

were collected before and after the onset of disease, we identified 53 epitopes that tracked with disease progression. For HSV, we identified epitopes in an envelope glycoprotein that were specific to either HSV1 or HSV2 and were corroborated by previous studies. We also identified 30 HSV2 epitopes that were 100% specific against HSV1. Several of these HSV2-specific epitopes were in previously described antigens. Notably, some of the HSV2-specific epitopes had sequences that were present in both the HSV1 and HSV2 proteomes. Thus, our analysis suggested that an epitope sequence that is present in two species' proteomes can still be specific for a single species. In analyzing Chagas disease specimens, we identified 222 epitopes in multiple novel and previously described antigens. Notably, these epitopes matched four out of seven peptides used in a Chagas diagnostic assay. By identifying epitopes in candidate autoantigens, we identified two Chagas-specific autoantigens that have also been described in an autoimmune disease with cardiac involvement. Additionally, we determined that several epitopes that appeared Chagas-specific, were also bound by leishmaniasis specimens. By subtracting out these epitopes, we identified epitopes that could differentiate between the two infections. With MESA, we determined the important binding positions and amino acid preferences in these disease-specifics epitopes. In one case, we determined binding motifs with MESA to show that two similar epitopes were bound by different antibodies.

Through this analysis, we demonstrated that K-TOPE and MESA can identify and characterize epitopes for a diverse set of diseases. We learned from these analyses that specimen collection has a strong bearing on the strength of a study. In the case of AMD, our disease and control specimens had multiple subtypes, which likely confounded analysis. Unfortunately, using a small sample size for HSV increases the uncertainty of whether the K-

TOPE epitopes will generalize to a larger population. Another observation from this study was that validating epitopes is complicated by the diversity of experimental and analytical approaches used by researchers. For epitopes that that cannot be validated by previous studies, it is difficult to know whether the epitopes are novel or spurious. Finally, the candidate antigens identified by this study may require further experimental validation to determine if they are involved in the progression of disease. These algorithms could be implemented by research groups that already have databases of antibody-binding peptides and could be applied to potentially hundreds of diseases with prominent antibody responses.

5.2 Future directions

Ultimately, these approaches were developed to aid the advancement of medical applications. Prior to these applications, the results generated by these tools will require further validation. To validate K-TOPE candidate antigens, the natively folded antigens could be spotted onto a microarray or ELISA [171] to assess whether the antigens bind sera. This validation would also enable the distinction between multiple candidate antigens. The antigens identified for AMD, HSV, and Chagas diseases would be suitable initial candidates for this type of validation. To further validate K-TOPE epitopes, we could compare disease-specific epitopes to motifs generated using alternative bioinformatic methods [43]. Once an antigen is confirmed as relevant to a disease, epitopes could be mapped on to the antigen's crystal structure [173]. The location of an epitope on an antigen could reveal the functional role of the corresponding antibody. For example, this analysis could identify a virus-neutralizing antibody [174]. Additionally, with MESA, we could determine if the binding motif of an epitope occurs at the antibody/antigen binding interface [17]. Thus, K-TOPE and

MESA could be used to discover antigens and characterize the antibody repertoire's response to these antigens.

To develop useful medical applications, it could be advantageous to analyze any of the multiple diseases have already been interrogated using peptide libraries. Examples of previously characterized diseases are Dengue virus [29], HIV [36], cancer [32], malaria [35], valley fever [20], and numerous other viral infections [28,34]. In these cases, any epitopes identified by K-TOPE could be validated using previous studies.

Downstream medical applications of K-TOPE and MESA include preventing, diagnosing, and treating diseases. A peptide containing multiple epitopes could be used as part of a vaccine formulation [101], conferring protection against multiple organisms. This would be a more focused approach than injecting a whole organism, in which case antibody generation is more stochastic. Also, K-TOPE enables feedback in vaccine formulation since it aids the process of correlating vaccine formulation with the epitopes targeted by the immune system. Thus, vaccine formulation could be altered to optimize the proportion of the population that binds to neutralizing epitopes. Diseases can be diagnosed by using a novel antigen in a conventional ELISA [172]. Alternatively, a panel of disease-specific epitopes could be printed on to a microarray as a diagnostic [46]. Then, machine learning could accurately assign disease status using microarray binding data [34]. Since treatments are often most effective in the early stages of disease, early detection of antibodies is crucial [30]. Diseases could be treated by developing monoclonal antibody therapeutics [157] using epitope and antigen information. These antibodies could be against conserved pathogen epitopes [42,145,151,152] and cancer [126,150]. MESA could aid therapeutic development by precisely characterizing epitope binding motifs. This would help avoid undesired cross

reactivity, which can be problematic for therapeutics [7,154]. Thus, K-TOPE and MESA could aid the development of a large variety of medical applications.

5.3 Overall conclusions

We developed two approaches, K-TOPE and MESA, which could greatly aid epitope identification and characterization. K-TOPE is designed to identify antigens and epitopes together, rather than in two separate steps. MESA can then computationally substitute these epitopes to identify the important binding positions and amino acid preferences. Importantly, both approaches use peptides selected from random libraries, enabling the identification and characterization of epitopes using a single experimental screen. We demonstrated the capabilities of these approaches by identifying disease-specific epitopes and antigens for age-related macular degeneration, herpes simplex virus, and Chagas disease. The results of these analyses suggest that these tools could be applicable to numerous additional diseases. Using the epitopes and antigens identified in this study, it could be possible to develop effective vaccines, diagnostics, and therapeutics. K-TOPE and MESA are novel tools for probing antibody repertoires that could significantly aid the development of medical applications.

6 References

1. Sun P, Ju H, Liu Z, Ning Q, Zhang J, Zhao X, Huang Y, Ma Z, Li Y: **Bioinformatics resources and tools for conformational B-cell epitope prediction**. *Comput Math Methods Med* 2013, **2013**, http://dx.doi.org/10.1155/2013/943636.

2. Paull ML, Daugherty PS: **Mapping serum antibody repertoires using peptide libraries**. *Curr Opin Chem Eng* 2018, **19**:21–26, http://dx.doi.org/10.1016/j.coche.2017.12.001.

3. Lemon SM, Gates NL, Simms TE, Bancroft WH: **IgM antibody to hepatitis B core antigen as a diagnostic parameter of acute infection with hepatitis B virus**. *J Infect Dis* 1981, **143**:803–809, http://dx.doi.org/10.1093/infdis/143.6.803.

4. Schellekens GA, Visser H, De Jong BA, Van Den Hoogen FH, Hazes JM, Breedveld FC, Van Venrooij WJ: **The diagnostic properties of rheumatoid arthritis antibodies recognizing a cyclic citrullinated peptide**. *Arthritis Rheum* 2000, **43**:155–163, http://dx.doi.org/10.1002/1529-0131(200001)43:1<155::AID-ANR20>3.0.CO;2-3.

5. Wide L, Bennich H, Johansson S: **Diagnosis of allergy by an in-vitro test for allergen antibodies**. *Lancet* 1967, **290**:1105–1107, http://dx.doi.org/10.1016/S0140-6736(67)90615-0.

6. Soria-Guerra RE, Nieto-Gomez R, Govea-Alonso DO, Rosales-Mendoza S: **An overview of bioinformatics tools for epitope prediction: Implications on vaccine development**. *J Biomed Inform* 2015, **53**:405–414, http://dx.doi.org/10.1016/j.jbi.2014.11.003.

7. Michaud GA, Salcius M, Zhou F, Bangham R, Bonin J, Guo H, Snyder M, Predki PF, Schweitzer BI: **Analyzing antibody specificity with whole proteome microarrays.** *Nat Biotechnol* 2003, **21**:1509–1512, http://dx.doi.org/10.1038/nbt910.

8. Forsström B, Axnäs BB, Stengele KP, Bühler J, Albert TJ, Richmond TA, Hu FJ, Nilsson P, Hudson EP, Rockberg J, et al.: **Proteome-wide epitope mapping of antibodies using ultra-dense peptide arrays**. *Mol Cell Proteomics* 2014, **13**:1585–1597, http://dx.doi.org/10.1074/mcp.M113.033308.

9. Stafford P, Wrapp D, Johnston SA: **General assessment of humoral activity in healthy humans.** *Mol Cell Proteomics* 2016, **15**:1610–1621, http://dx.doi.org/10.1074/mcp.M115.054601.

10. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR: **The promise and challenge of high-throughput sequencing of the antibody repertoire.** *Nat Biotechnol* 2014, **32**:158–68, http://dx.doi.org/10.1038/nbt.2782.

11. Lavinder JJ, Horton AP, Georgiou G, Ippolito GC: **Next-generation sequencing and protein mass spectrometry for the comprehensive analysis of human cellular and serum antibody repertoires**. *Curr Opin Chem Biol* 2015, **24**:112–120,

http://dx.doi.org/10.1016/j.cbpa.2014.11.007.

12. Horns F, Vollmers C, Croote D, Mackey SF, Swan GE, Dekker CL, Davis MM, Quake SR: **Lineage tracing of human B cells reveals the in vivo landscape of human antibody class switching**. *Elife* 2016, **5**:1–20, http://dx.doi.org/10.7554/eLife.16578.001.

13. Potocnakova L, Bhide M, Pulzova LB: **An introduction to B-cell epitope mapping and in silico epitope prediction**. *J Immunol Res* 2016, **2016**:1–11, http://dx.doi.org/10.1155/2016/6760830.

14. Yao B, Zheng D, Liang S, Zhang C: **Conformational B-cell epitope prediction on antigen protein structures: A review of current algorithms and comparison with common binding site prediction methods**. *PLoS One* 2013, **8**:22–25, http://dx.doi.org/10.1371/journal.pone.0062249.

15. Van Regenmortel MHV: **Mapping epitope structure and activity: From one-dimensional prediction to four-dimensional description of antigenic specificity**. *Methods a Companion To Methods Enzymol* 1996, **9**:465–472, http://dx.doi.org/10.1006/meth.1996.0054.

16. Sun J, Xu T, Wang S, Li G, Wu D, Cao Z: **Does difference exist between epitope and non-epitope residues? Analysis of the physicochemical and structural properties on conformational epitopes from B-cell protein antigens**. *Immunome Res* 2011, **7**:1–11.

17. Kringelum JV, Nielsen M, Padkjær SB, Lund O: **Structural analysis of B-cell epitopes in antibody:protein complexes**. *Mol Immunol* 2013, **53**:24–34, http://dx.doi.org/10.1016/j.molimm.2012.06.001.

18. Sivalingam GN, Shepherd AJ: **An analysis of B-cell epitope discontinuity**. *Mol Immunol* 2012, **51**:304–309, http://dx.doi.org/10.1016/j.molimm.2012.03.030.

19. Pashova S, Schneider C, Gunten S von, Pashov A: **Antibody repertoire profiling with mimotope arrays**. *Hum Vaccines Immunother* 2017, **13**:314–322, http://dx.doi.org/10.1080/21645515.2017.1264786.

20. Navalkar KA, Johnston SA, Stafford P: **Peptide based diagnostics: Are random-sequence peptides more useful than tiling proteome sequences?** *J Immunol Methods* 2015, **417**:10–21, http://dx.doi.org/10.1016/j.jim.2014.12.002.

21. Larman HB, Laserson U, Querol L, Verhaeghen K, Solimini NL, Xu GJ, Klarenbeek PL, Church GM, Hafler DA, Plenge RM, et al.: **PhIP-Seq characterization of autoantibodies from patients with multiple sclerosis, type 1 diabetes and rheumatoid arthritis**. *J Autoimmun* 2013, **43**:1–9, http://dx.doi.org/10.1016/j.jaut.2013.01.013.

22. Frietze KM, Roden RBS, Lee J-H, Shi Y, Peabody DS, Chackerian B: **Identification of anti-CA125 antibody responses in ovarian cancer patients by a novel deep sequence-coupled biopanning platform**. *Cancer Immunol Res* 2016, **4**:157–164,

http://dx.doi.org/10.1158/2326-6066.CIR-15-0165.

23. Hecker M, Fitzner B, Wendt M, Lorenz P, Flechtner K, Steinbeck F, Schröder I, Thiesen HJ, Zettl UK: **High-density peptide microarray analysis of IgG autoantibody reactivities in serum and cerebrospinal fluid of multiple sclerosis patients.** *Mol Cell Proteomics* 2016, **15**:1360–80, http://dx.doi.org/10.1074/mcp.M115.051664.

24. Carmona SJ, Nielsen M, Schafer-Nielsen C, Mucci J, Altcheh J, Balouz V, Tekiel V, Frasch AC, Campetella O, Buscaglia CA, et al.: **Towards high-throughput immunomics for infectious diseases: Use of next-generation peptide microarrays for rapid discovery and mapping of antigenic determinants**. *Mol Cell Proteomics* 2015, **14**:1871–1884, http://dx.doi.org/10.1074/mcp.M114.045906.

25. Zandian A, Forsström B, Häggmark-Manberg A, Schwenk JM, Uhlén M, Nilsson P, Ayoglu B: **Whole-proteome peptide microarrays for profiling autoantibody repertoires within multiple sclerosis and narcolepsy.** *J Proteome Res* 2017, **16**:1300–1314, http://dx.doi.org/10.1021/acs.jproteome.6b00916.

26. Buus S, Rockberg J, Forsström B, Nilsson P, Uhlen M, Schafer-Nielsen C: **High-resolution mapping of linear antibody epitopes using ultrahigh-density peptide microarrays.** *Mol Cell Proteomics* 2012, **11**:1790–1800, http://dx.doi.org/10.1074/mcp.M112.020800.

27. Larman HB, Zhao Z, Laserson U, Li MZ, Ciccia A, Gakidis MAM, Church GM, Kesari S, LeProust EM, Solimini NL, et al.: **Autoantigen discovery with a synthetic human peptidome**. *Nat Biotechnol* 2011, **29**:535–541, http://dx.doi.org/10.1038/nbt.1856.

28. Xu GJ, Kula T, Xu Q, Li MZ, Vernon SD, Ndung'u T, Ruxrungtham K, Sanchez J, Brander C, Chung RT, et al.: **Comprehensive serological profiling of human populations using a synthetic human virome.** *Science (80-)* 2015, **348**:aaa0698, http://dx.doi.org/10.1126/science.aaa0698.

29. Frietze KM, Pascale JM, Moreno B, Chackerian B, Peabody DS: **Pathogen-specific deep sequence-coupled biopanning: A method for surveying human antibody responses**. *PLoS One* 2017, **12**:e0171511, http://dx.doi.org/10.1371/journal.pone.0171511.

30. Restrepo L, Stafford P, Johnston SA, Albert S: **Feasibility of an early Alzheimer's disease immunosignature diagnostic test**. *J Neuroimmunol* 2012, **254**:154–160, http://dx.doi.org/10.1016/j.jneuroim.2012.09.014.

31. Kukreja M, Johnston SA, Stafford P: **Immunosignaturing microarrays distinguish antibody profiles of related pancreatic diseases**. *J Proteomics Bioinform* 2012, **S6**:1–5, http://dx.doi.org/10.4172/jpb.S6-001.

32. Legutki JB, Zhao Z-G, Greving M, Woodbury N, Johnston SA, Stafford P: **Scalable high-density peptide arrays for comprehensive health monitoring.** *Nat Commun* 2014, **5**:4785, http://dx.doi.org/10.1038/ncomms5785.

33. Singh S, Stafford P, Schlauch KA, Tillett RR, Gollery M, Johnston SA, Khaiboullina SF, De Meirleir KL, Rawat S, Mijatovic T, et al.: **Humoral immunity profiling of subjects with myalgic encephalomyelitis using a random peptide microarray differentiates cases from controls with high specificity and sensitivity**. *Mol Neurobiol* 2016, **2016**:1–9, http://dx.doi.org/10.1007/s12035-016-0334-0.

34. Stafford P, Cichacz Z, Woodbury NW, Johnston SA: **Immunosignature system for diagnosis of cancer**. *Proc Natl Acad Sci* 2014, **111**:E3072–E3080, http://dx.doi.org/10.1073/pnas.1409432111.

35. Richer J, Johnston SA, Stafford P: **Epitope identification from fixed-complexity random-sequence peptide microarrays**. *Mol Cell Proteomics* 2015, **14**:136–147, http://dx.doi.org/10.1074/mcp.M114.043513.

36. Ryvkin A, Ashkenazy H, Smelyanski L, Kaplan G, Penn O, Weiss-Ottolenghi Y, Privman E, Ngam PB, Woodward JE, May GD, et al.: **Deep panning: Steps towards probing the IgOme**. *PLoS One* 2012, **7**:1–11, http://dx.doi.org/10.1371/journal.pone.0041469.

37. Bachler BC, Humbert M, Palikuqi B, Siddappa NB, Lakhashe SK, Rasmussen RA, Ruprecht RM: **Novel biopanning strategy to identify epitopes associated with vaccine protection**. *J Virol* 2013, **87**:4403–4416, http://dx.doi.org/10.1128/JVI.02888-12.

38. Liu X, Hu Q, Liu S, Tallo LJ, Sadzewicz L, Schettine CA, Nikiforov M, Klyushnenkova EN, Ionov Y: **Serum antibody repertoire profiling using in silico antigen screen**. *PLoS One* 2013, **8**:e67181, http://dx.doi.org/10.1371/journal.pone.0067181.

39. Weber LK, Palermo A, Kügler J, Armant O, Isse A, Rentschler S, Jaenisch T, Duebel S, Nesterov-Mueller A, Loeffler FF: **Single amino acid fingerprinting of the human antibody repertoire with high density peptide arrays**. *J Immunol Methods* 2017, **443**:45–54, http://dx.doi.org/10.1016/j.jim.2017.01.012.

40. Christiansen A, Kringelum J V, Hansen CS, Bøgh KL, Sullivan E, Patel J, Rigby NM, Eiwegger T, Szépfalusi Z, Masi F De, et al.: **High-throughput sequencing enhanced phage display enables the identification of patient-specific epitope motifs in serum**. *Sci Rep* 2015, **5**:1–13, http://dx.doi.org/10.1038/srep12913.

41. Spatola BN, Murray JA, Kagno M, Kaukinen K, Daugherty PS: **Antibody repertoire profiling using bacterial display identifies reactivity signatures of celiac disease**. *Anal Chem* 2012, **85**:1215–1222, http://dx.doi.org/10.1021/ac303201d.

42. Elliott SE, Parchim NF, Kellems RE, Xia Y, Soffici AR, Daugherty PS: **A pre-eclampsia-associated Epstein-Barr virus antibody cross-reacts with placental GPR50**. *Clin Immunol* 2016, **168**:64–71, http://dx.doi.org/10.1016/j.clim.2016.05.002.

43. Pantazes RJ, Reifert J, Bozekowski J, Ibsen KN, Murray JA, Daugherty PS: **Identification of disease-specific motifs in the antibody specificity repertoire via next-generation sequencing**. *Sci Rep* 2016, **6**:30312,

http://dx.doi.org/10.1038/srep30312.

44. Hansen CS, Østerbye T, Marcatili P, Lund O, Buus S, Nielsen M: **ArrayPitope : Automated analysis of amino acid substitutions for peptide microarray-based antibody epitope mapping**. *PLoS One* 2017, **278832**:1–14, http://dx.doi.org/10.1371/journal.pone.0168453.

45. Burnham CAD, McAdam AJ: **Your viral past: A comprehensive method for serological profiling to explore the human virome**. *Clin Chem* 2016, **62**:426–427, http://dx.doi.org/10.1373/clinchem.2015.245027.

46. Sykes KF, Legutki JB, Stafford P: **Immunosignaturing: A critical review**. *Trends Biotechnol* 2013, **31**:45–51, http://dx.doi.org/10.1016/j.tibtech.2012.10.012.

47. Halperin RF, Stafford P, Johnston SA: **Exploring antibody recognition of sequence space through random-sequence peptide microarrays.** *Mol Cell Proteomics* 2011, **10**:M110.000786, http://dx.doi.org/10.1074/mcp.M110.000786.

48. Stafford P, Halperin R, Legutki JB, Magee DM, Galgiani J, Johnston SA: **Physical characterization of the "immunosignaturing effect."** *Mol Cell Proteomics* 2012, **11**:M111.011593-M111.011593, http://dx.doi.org/10.1074/mcp.M111.011593.

49. Bailey T, Elkan C: **Unsupervised learning of multiple motifs using expected minimization**. *Mach Learn* 1995, **21**:51–80, http://dx.doi.org/10.1007/BF00993379.

50. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403–10, http://dx.doi.org/10.1016/S0022-2836(05)80360-2.

51. Bastas G, Sompuram SR, Pierce B, Vani K, Bogen SA: **Bioinformatic requirements for protein database searching using predicted epitopes from disease-associated antibodies**. *Mol Cell Proteomics* 2007, **7**:247–256, http://dx.doi.org/10.1074/mcp.M700107-MCP200.

52. Rockberg J, Lofblom J, Hjelm B, Uhlen M, Stahl S: **Epitope mapping of antibodies using bacterial surface display**. *Nat Methods* 2008, **5**:1039–1045, http://dx.doi.org/10.1038/nmeth.1272.

53. Daugherty PS: **Protein engineering with bacterial display**. *Curr Opin Struct Biol* 2007, **17**:474–480, http://dx.doi.org/10.1016/j.sbi.2007.07.004.

54. Kim T, Tyndel MS, Huang H, Sidhu SS, Bader GD, Gfeller D, Kim PM: **MUSI: An integrated system for identifying multiple specificity from very large peptide or nucleic acid data sets**. *Nucleic Acids Res* 2012, **40**:e47, http://dx.doi.org/10.1093/nar/gkr1294.

55. Andreatta M, Lund O, Nielsen M: **Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach**. *Bioinformatics* 2013, **29**:8–14, http://dx.doi.org/10.1093/bioinformatics/bts621.

56. Liu H, Han F, Zhou H, Yan X, Kosik KS: **Fast motif discovery in short sequences**.

2016 IEEE 32nd Int Conf Data Eng ICDE 2016 2016, **2016**:1158–1169, http://dx.doi.org/10.1109/ICDE.2016.7498321.

57. Bressler NM: **Age-related macular degeneration is the leading cause of blindness**. *JAMA* 2004, **291**:1900–1901, http://dx.doi.org/10.1093/brain/aws003.

58. Castle SC: **Clinical relevance of age-related immune dysfunction**. *Clin Infect Dis* 2000, **31**:578–585, http://dx.doi.org/1058-4838/2000.

59. Lambert NG, ElShelmani H, Singh MK, Mansergh FC, Wride MA, Padilla M, Keegan D, Hogg RE, Ambati BK: **Risk factors and biomarkers of age-related macular degeneration**. *Prog Retin Eye Res* 2016, **54**:64–102, http://dx.doi.org/10.1016/j.preteyeres.2016.04.003.

60. Gehrs KM, Anderson DH, Johnson L V, Hageman GS: **Age-related macular degeneration - Emerging pathogenetic and therapeutic concepts**. *Ann Med* 2006, **38**:450–471, http://dx.doi.org/10.1080/07853890600946724.

61. Flaxman SR, Bourne RRA, Resnikoff S, Ackland P, Braithwaite T, Cicinelli M V., Das A, Jonas JB, Keeffe J, Kempen JH, et al.: **Global causes of blindness and distance vision impairment 1990–2020: A systematic review and meta-analysis**. *Lancet Glob Heal* 2017, **5**:e1221–e1234, http://dx.doi.org/10.1016/S2214-109X(17)30393-5.

62. Wong WL, Su X, Li X, Cheung CMG, Klein R, Cheng CY, Wong TY: **Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: A systematic review and meta-analysis**. *Lancet Glob Heal* 2014, **2**:e106–e116, http://dx.doi.org/10.1016/S2214-109X(13)70145-1.

63. Hollyfield JG: **Age-related macular degeneration: The molecular link between oxidative damage, tissue-specific inflammation and outer retinal disease**. *Investig Ophthalmol Vis Sci* 2010, **51**:1276–1281, http://dx.doi.org/10.1167/iovs.09-4478.

64. Strauss O: **The retinal pigment epithelium in visual function**. *Physiol Rev* 2005, **85**:845–881, http://dx.doi.org/10.1152/physrev.00021.2004.

65. Mullins RF, Russell SR, Anderson DH, Hageman GS: **Drusen associated with aging and age-related macular degeneration contain proteins common to extracellular deposits associated with atherosclerosis, elastosis, amyloidosis, and dense deposit disease.** *FASEB J* 2000, **14**:835–846, http://dx.doi.org/10.1096/fasebj.14.7.835.

66. Crabb JW, Miyagi M, Gu X, Shadrach K, West KA, Sakaguchi H, Kamei M, Hasan A, Yan L, Rayborn ME, et al.: **Drusen proteome analysis: An approach to the etiology of age-related macular degeneration.** *Proc Natl Acad Sci U S A* 2002, **99**:14682–7, http://dx.doi.org/10.1073/pnas.222551899.

67. Liu M, Regillo CD: **A review of treatments for macular degeneration: A synopsis of currently approved treatments and ongoing clinical trials**. *Curr Opin Ophthalmol* 2004, **15**:221–226, http://dx.doi.org/10.1097/01.icu.0000122122.24016.f1.

68. Donoso LA, Kim D, Frost A, Callahan A, Hageman G: **The role of inflammation in the pathogenesis of age-related macular degeneration**. *Surv Ophthalmol* 2006, **51**:137–152, http://dx.doi.org/10.1016/j.survophthal.2005.12.001.

69. Kaarniranta K, Salminen A: **Age-related macular degeneration: Activation of innate immunity system via pattern recognition receptors**. *J Mol Med* 2009, **87**:117–123, http://dx.doi.org/10.1007/s00109-008-0418-z.

70. Whitcup SM, Sodhi A, Atkinson JP, Holers VM, Sinha D, Rohrer B, Dick AD, Wilmer T, Johns T: **The role of the immune response in age-related macular degeneration**. *Int J Inflam* 2013, **2013**, http://dx.doi.org/10.1155/2013/348092 Review.

71. Camelo S: **Potential sources and roles of adaptive immunity in age-related macular degeneration: Shall we rename AMD into autoimmune macular disease?** *Autoimmune Dis* 2014, **2014**:27–29, http://dx.doi.org/10.1155/2014/532487.

72. Thakkinstian A, Han P, McEvoy M, Smith W, Hoh J, Magnusson K, Zhang K, Attia J: **Systematic review and meta-analysis of the association between complementary factor H Y402H polymorphisms and age-related macular degeneration**. *Hum Mol Genet* 2006, **15**:2784–2790, http://dx.doi.org/10.1093/hmg/ddl220.

73. Le KN, Gibiansky L, Van Lookeren Campagne M, Good J, Davancaze T, Loyet KM, Morimoto A, Strauss EC, Jin JY: **Population pharmacokinetics and pharmacodynamics of Lampalizumab administered intravitreally to patients with geographic atrophy**. *CPT Pharmacometrics Syst Pharmacol* 2015, **4**:595–604, http://dx.doi.org/10.1002/psp4.12031.

74. Penfold PL, Provis JM, Furby JH, Gatenby PA, Billson FA: **Autoantibodies to retinal astrocyte associated with age-related macular degeneration**. *Graefe's Arch Clin Exp Opthalmology* 1990, **228**:270–274, http://dx.doi.org/10.1007/BF00920033.

75. Umeda S, Suzuki MT, Okamoto H, Ono F, Mizota A, Terao K, Yoshikawa Y, Tanaka Y, Iwata T: **Molecular composition of drusen and possible involvement of anti-retinal autoimmunity in two different forms of macular degeneration in cynomolgus monkey (Macaca fascicularis)**. *FASEB J* 2005, **19**:1683–1685, http://dx.doi.org/10.1096/fj.04-3525fje.

76. Patel N, Ohbayashi M, Nugent AK, Ramchand K, Toda M, Chau KY, Bunce C, Webster A, Bird AC, Ono SJ, et al.: **Circulating anti-retinal antibodies as immune markers in age-related macular degeneration**. *Immunology* 2005, **115**:422–430, http://dx.doi.org/10.1111/j.1365-2567.2005.02173.x.

77. Joachim SC, Bruns K, Lackner KJ, Pfeiffer N, Grus FH: **Analysis of IgG antibody patterns against retinal antigens and antibodies to α-crystallin, GFAP, and α-enolase in sera of patients with "wet" age-related macular degeneration**. *Graefe's Arch Clin Exp Ophthalmol* 2007, **245**:619–626, http://dx.doi.org/10.1007/s00417-006-0429-9.

78. Iannaccone A, Giorgianni F, New DD, Hollingsworth TJ, Umfress A, Alhatem AH,

Neeli I, Lenchik NI, Jennings BJ, Calzada JI, et al.: **Circulating autoantibodies in age-related macular degeneration recognize human macular tissue antigens implicated in autophagy, immunomodulation, and protection from oxidative stress and apoptosis**. *PLoS One* 2015, **10**:1–22, http://dx.doi.org/10.1371/journal.pone.0145323.

79. Adamus G, Chew EY, Ferris FL, Klein ML: **Prevalence of anti-retinal autoantibodies in different stages of age-related macular degeneration**. *BMC Ophthalmol* 2014, **14**:1–9, http://dx.doi.org/10.1186/1471-2415-14-154.

80. Morohoshi K, Patel N, Ohbayashi M, Chong V, Grossniklaus HE, Bird AC, Ono SJ: **Serum autoantibody biomarkers for age-related macular degeneration and possible regulators of neovascularization**. *Exp Mol Pathol* 2011, **92**:64–73, http://dx.doi.org/10.1016/j.yexmp.2011.09.017.

81. Morohoshi K, Ohbayashi M, Patel N, Chong V, Bird AC, Ono SJ: **Identification of anti-retinal antibodies in patients with age-related macular degeneration**. *Exp Mol Pathol* 2012, **93**:193–199, http://dx.doi.org/10.1016/j.yexmp.2012.03.007.

82. Adamus G: **Can innate and autoimmune reactivity forecast early and advance stages of age-related macular degeneration?** *Autoimmun Rev* 2017, **16**:231–236, http://dx.doi.org/10.1016/j.autrev.2017.01.005.

83. Taylor TJ, Brockman MA, Mcnamee EE, Knipe DM: **Herpes simplex virus**. *Front Biosci* 2002, **7**:752–764, http://dx.doi.org/10.2741/taylor.

84. Whitley RJ, Roizman B: **Herpes simplex virus infections**. *Lancet* 2001, **357**:1513–1518, http://dx.doi.org/10.1016/S0140-6736(00)04638-9.

85. Looker KJ, Magaret AS, May MT, Turner KME, Vickerman P, Gottlieb SL, Newman LM: **Global and regional estimates of prevalent and incident herpes simplex virus type 1 infections in 2012**. *PLoS One* 2015, **10**:1–17, http://dx.doi.org/10.1371/journal.pone.0140765.

86. Looker KJ, Magaret AS, Turner KME, Vickerman P, Gottlieb SL, Newman LM: **Global estimates of prevalent and incident herpes simplex virus type 2 infections in 2012**. *PLoS One* 2015, **10**:1–23, http://dx.doi.org/10.1371/journal.pone.0114989.

87. Bowden RJ, Mcgeoch DJ: *Evolution of herpes simplex viruses*. CRC Press; 2017.

88. Gupta R, Warren T, Wald A: **Genital herpes**. *Lancet* 2007, **370**:2127–2137, http://dx.doi.org/10.1016/S0140-6736(07)61908-4.

89. McGeoch DJ, Moss HWM, McNab D, Frame MC: **DNA sequence and genetic content of the HindIII l region in the short unique component of the herpes simplex virus type 2 genome: Identification of the gene encoding glycoprotein G, and evolutionary comparisons**. *J Gen Virol* 1987, **68**:19–38, http://dx.doi.org/10.1099/0022-1317-68-1-19.

90. Marsden HS, MacAulay K, Murray J, Smith IW: **Identification of an immunodominant sequential epitope in glycoprotein G of herpes simplex virus**

type 2 that is useful for serotype-specific diagnosis. *J Med Virol* 1998, **56**:79–84, http://dx.doi.org/10.1002/(SICI)1096-9071(199809)56:1<79::AID-JMV13>3.0.CO;2-R.

91. Liljeqvist JÅ, Trybala E, Svennerholm B, Jeansson S, Sjögren-Jansson E, Bergström T: **Localization of type-specific epitopes of herpes simplex virus type 2 glycoprotein G recognized by human and mouse antibodies**. *J Gen Virol* 1998, **79**:1215–1224, http://dx.doi.org/10.1099/0022-1317-79-5-1215.

92. Grabowska A, Jameson C, Laing P, Jeansson S, Sjögren-Jansson E, Taylor J, Cunningham A, Irving WL: **Identification of type-specific domains within glycoprotein G of herpes simplex virus type 2 (HSV-2) recognized by the majority of patients infected with HSV-2, but not by those infected with HSV-1**. *J Gen Virol* 1999, **80**:1789–1798, http://dx.doi.org/10.1099/0022-1317-80-7-1789.

93. Oladepo DK, Klapper PE, Marsden HS: **Peptide based enzyme-linked immunoassays for detection of anti-HSV-2 IgG in human sera**. *J Virol Methods* 2000, **87**:63–70, http://dx.doi.org/10.1016/S0166-0934(00)00152-X.

94. Nilsen A, Ulvestad E, Marsden H, Langeland N, Myrmel H, Matre R, Haarr L: **Performance characteristics of a glycoprotein G based oligopeptide (peptide 55) and two different methods using the complete glycoprotein as assays for detection of anti-HSV-2 antibodies in human sera**. *J Virol Methods* 2003, **107**:21–27, http://dx.doi.org/10.1016/S0166-0934(02)00185-4.

95. Cairns TM, Shaner MS, Zuo Y, Baribaud I, Eisenberg RJ, Cohen GH, Whitbeck JC, Ponce-de-leon M: **Epitope mapping of herpes simplex virus type 2 gH / gL defines distinct antigenic sites, including some associated with biological function**. *J Virol* 2006, **80**:2596–2608, http://dx.doi.org/10.1128/JVI.80.6.2596.

96. Clo E, Kracun SK, Nudelman AS, Jensen KJ, Liljeqvist J-A, Olofsson S, Bergstrom T, Blixt O: **Characterization of the viral O-glycopeptidome: A novel tool of relevance for vaccine design and serodiagnosis**. *J Virol* 2012, **86**:6268–6278, http://dx.doi.org/10.1128/JVI.00392-12.

97. Pan M, Wang X, Liao J, Yin D, Li S, Pan Y, Wang Y, Xie G, Zhang S, Li Y: **Prediction and identification of potential immunodominant epitopes in glycoproteins B, C, E, G, and i of herpes simplex virus type 2**. *Clin Dev Immunol* 2012, **2012**, http://dx.doi.org/10.1155/2012/205313.

98. Liu K, Jiang D, Zhang L, Yao Z, Chen Z, Yu S, Wang X: **Identification of B- and T-cell epitopes from glycoprotein B of herpes simplex virus 2 and evaluation of their immunogenicity and protection efficacy**. *Vaccine* 2012, **30**:3034–3041, http://dx.doi.org/10.1016/j.vaccine.2011.10.010.

99. Whitbeck JC, Huang Z-Y, Cairns TM, Gallagher JR, Lou H, Ponce-de-Leon M, Belshe RB, Eisenberg RJ, Cohen GH: **Repertoire of epitopes recognized by serum IgG from humans vaccinated with herpes simplex virus 2 glycoprotein D**. *J Virol* 2014, **88**:7786–7795, http://dx.doi.org/10.1128/JVI.00544-14.

100. Ackermann G, Ackermann F, Eggers HJ, Wieland U, Kühn JE: **Mapping of linear antigenic determinants on glycoprotein C of herpes simplex virus type 1 and type 2 recognized by human serum immunoglobulin G antibodies.** *J Med Virol* 1998, **55**:281–7, http://dx.doi.org/10.1002/(SICI)1096-9071(199808)55:4<281::AID-JMV5>3.0.CO;2-X.

101. Risinger C, Sørensen KK, Jensen KJ, Olofsson S, Bergström T, Blixt O: **Linear multiepitope (glyco)peptides for type-specific serology of herpes simplex virus (HSV) infections**. *ACS Infect Dis* 2017, **3**:360–367, http://dx.doi.org/10.1021/acsinfecdis.7b00001.

102. Rassi A, Rassi A, Marin-Neto JA: **Chagas disease**. *Lancet* 2010, **375**:1388–1402, http://dx.doi.org/10.1016/S0140-6736(10)60061-X.

103. Coura JR, Junqueira ACV, Fernandes O, Valente SAS, Miles MA: **Emerging Chagas disease in Amazonian Brazil**. *Trends Parasitol* 2002, **18**:171–176, http://dx.doi.org/10.1016/S1471-4922(01)02200-0.

104. Scares MBP, Pontes-De-Carvalho L, Ribeiro-Dos-Santos R: **The pathogenesis of Chagas' disease: When autoimmune and parasite-specific immune responses meet**. *An Acad Bras Cienc* 2001, **73**:546–559, http://dx.doi.org/10.1590/S0001-37652001000400008.

105. Bern C: **Chagas' disease**. *N Engl J Med* 2015, **373**:456–466, http://dx.doi.org/10.1056/NEJMra1410150.

106. Moncayo Á, Silveira AC: **Current epidemiological trends of Chagas disease in Latin America and future challenges: Epidemiology, surveillance, and health policies**. *Am Trypanos Chagas Dis One Hundred Years Res Second Ed* 2017, **104**:59–88, http://dx.doi.org/10.1016/B978-0-12-801029-7.00004-6.

107. Afonso AM, Ebell MH, Tarleton RL: **A systematic review of high quality diagnostic tests for Chagas disease**. *PLoS Negl Trop Dis* 2012, **6**, http://dx.doi.org/10.1371/journal.pntd.0001881.

108. Balouz V, Agüero F, Buscaglia CA: **Chagas disease diagnostic applications**. *Adv Parasitol* 2017, **97**:1–45, http://dx.doi.org/10.1016/bs.apar.2016.10.001.

109. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran A-N, Ghedin E, Worthey EA, Delcher AL, Blandin G, et al.: **The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease**. *Science (80-)* 2005, **309**:409–15, http://dx.doi.org/10.1126/science.1112631.

110. Balouz V, De Cámara MLM, Cánepa GE, Carmona SJ, Volcovich R, Gonzalez N, Altcheh J, Agüero F, Buscaglia CA: **Mapping antigenic motifs in the trypomastigote small surface antigen from Trypanosoma cruzi**. *Clin Vaccine Immunol* 2015, **22**:304–312, http://dx.doi.org/10.1128/CVI.00684-14.

111. Di Noia JM, Buscaglia CA, De Marchi CR, Almeida IC, Frasch ACC: **A Trypanosoma cruzi small surface molecule provides the first immunological**

evidence that Chagas' disease is due to a single parasite lineage. *J Exp Med* 2002, **195**:401–413, http://dx.doi.org/10.1084/jem.20011433.

112. de los Milagros Cámara M, Cánepa GE, Lantos AB, Balouz V, Yu H, Chen X, Campetella O, Mucci J, Buscaglia CA: **The Trypomastigote Small Surface Antigen (TSSA) regulates Trypanosoma cruzi infectivity and differentiation**. *PLoS Negl Trop Dis* 2017, **11**:1–21, http://dx.doi.org/10.1371/journal.pntd.0005856.

113. Balouz V, Melli LJ, Volcovich R, Moscatelli G, Moroni S, González N, Ballering G, Bisio M, Ciocchini AE, Buscaglia CA, et al.: **The trypomastigote small surface antigen from Trypanosoma cruzi improves treatment evaluation and diagnosis in pediatric chagas disease**. *J Clin Microbiol* 2017, **55**:3444–3453, http://dx.doi.org/10.1128/JCM.01317-17.

114. Rowe M, Melnick J, Gerwien R, Legutki JB, Pfeilsticker J, Tarasow TM, Sykes KF: **An immunosignature test distinguishes Trypanosoma cruzi, hepatitis B, hepatitis C and West Nile virus seropositivity among asymptomatic blood donors**. *PLoS Negl Trop Dis* 2017, **11**:e0005882, http://dx.doi.org/10.1371/journal.pntd.0005882.

115. Daugherty P, Kamath K, Reifert J: *Methods and compositions for assessing antibody specificities*. U.S. Patent Application No. WO2017083874A1; 2016.

116. Mucci J, Carmona SJ, Volcovich R, Altcheh J, Bracamonte E, Marco JD, Nielsen M, Buscaglia CA, Agüero F: **Next-generation ELISA diagnostic assay for Chagas disease based on the combination of short peptidic epitopes**. *PLoS Negl Trop Dis* 2017, **11**:1–19, http://dx.doi.org/10.1371/journal.pntd.0005972.

117. Pitcovsky TA, Buscaglia CA, Mucci J, Campetella O: **A functional network of intramolecular cross-reacting epitopes delays the elicitation of neutralizing antibodies to Trypanosoma cruzi trans-sialidase.** *J Infect Dis* 2002, **186**:397–404, http://dx.doi.org/10.1086/341463.

118. Lander N, Bernal C, Diez N, Añez N, Docampo R, Ramírez JL: **Localization and developmental regulation of a dispersed gene family 1 protein in Trypanosoma cruzi**. *Infect Immun* 2010, **78**:231–240, http://dx.doi.org/10.1128/IAI.00780-09.

119. Umezawa ES, Bastos SF, Coura JR, Levin MJ, Gonzalez A, Rangel-Aldao R, Zingales B, Luquetti AO, Da Silveira JF: **An improved serodiagnostic test for Chagas' disease employing a mixture of Trypanosoma cruzi recombinant antigens**. *Transfusion* 2003, **43**:91–97, http://dx.doi.org/10.1046/j.1537-2995.2003.00279.x.

120. Durante IM, La Spina PE, Carmona SJ, Agüero F, Buscaglia CA: **High-resolution profiling of linear B-cell epitopes from mucin-associated surface proteins (MASPs) of Trypanosoma cruzi during human infections**. *PLoS Negl Trop Dis* 2017, **11**:1–23, http://dx.doi.org/10.1371/journal.pntd.0005986.

121. dos Santos SL, Freitas LM, Lobo FP, Rodrigues-Luiz GF, Mendes TA de O, Oliveira ACS, Andrade LO, Chiari É, Gazzinelli RT, Teixeira SMR, et al.: **The MASP family of Trypanosoma cruzi: Changes in gene expression and antigenic profile during the acute phase of experimental infection**. *PLoS Negl Trop Dis* 2012, **6**,

http://dx.doi.org/10.1371/journal.pntd.0001779.

122. Bartholomeu DC, Cerqueira GC, Leão ACA, daRocha WD, Pais FS, Macedo C, Djikeng A, Teixeira SMR, El-Sayed NM: **Genomic organization and expression profile of the mucin-associated surface protein (masp) family of the human pathogen Trypanosoma cruzi**. *Nucleic Acids Res* 2009, **37**:3407–3417, http://dx.doi.org/10.1093/nar/gkp172.

123. Caballero ZC, Sousa OE, Marques WP, Saez-Alquezar A, Umezawa ES: **Evaluation of serological tests to identify Trypanosoma cruzi infection in humans and determine cross-reactivity with Trypanosoma rangeli and Leishmania spp.** *Clin Vaccine Immunol* 2007, **14**:1045–1049, http://dx.doi.org/10.1128/CVI.00127-07.

124. Ibsen KN, Daugherty PS: **Prediction of antibody structural epitopes via random peptide library screening and next generation sequencing**. *J Immunol Methods* 2017, **451**:28–36, http://dx.doi.org/10.1016/j.jim.2017.08.004.

125. Amanna IJ, Carlson NE, Slifka MK: **Duration of humoral immunity to common viral and vaccine antigens.** *N Engl J Med* 2007, **357**:1903–1915, http://dx.doi.org/10.1542/peds.2008-2139LLLL.

126. Gubin MM, Zhang X, Schuster H, Caron E, Ward JP, Noguchi T, Ivanova Y, Hundal J, Arthur CD, Krebber WJ, et al.: **Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens**. *Nature* 2014, **515**:577–581, http://dx.doi.org/10.1038/nature13988.

127. Lee J, Boutz DR, Chromikova V, Joyce MG, Vollmers C, Leung K, Horton AP, DeKosky BJ, Lee C-H, Lavinder JJ, et al.: **Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination**. *Nat Med* 2016, **22**:1456–1464, http://dx.doi.org/10.1038/nm.4224.

128. Sundström P, Nyström M, Ruuth K, Lundgren E: **Antibodies to specific EBNA-1 domains and HLA DRB1*1501 interact as risk factors for multiple sclerosis**. *J Neuroimmunol* 2009, **215**:102–107, http://dx.doi.org/10.1016/j.jneuroim.2009.08.004.

129. Samuelson A, Forsgren M, Johansson BO, Wahren B: **Molecular basis for serological cross-reactivity between enteroviruses**. *Clin Diagn Lab Immunol* 1994, **1**:336–341.

130. Young LS, Rickinson AB: **Epstein-Barr virus: 40 years on**. *Nat Rev Cancer* 2004, **4**:757–768, http://dx.doi.org/10.1038/nrc1452.

131. Romero Pastrana F, Neef J, Koedijk DGAM, de Graaf D, Duipmans J, Jonkman MF, Engelmann S, van Dijl JM, Buist G: **Human antibody responses against non-covalently cell wall-bound Staphylococcus aureus proteins.** *Sci Rep* 2018, **8**:3234, http://dx.doi.org/10.1038/s41598-018-21724-z.

132. Kaplan EL, Huwe BB: **The sensitivity and specificity of an agglutination test for antibodies to streptococcal extracellular antigens: A quantitative analysis and comparison of the streptozyme test with the anti-streptolysin O and anti-**

deoxyribonuclease B tests. *J Pediatr* 1980, **96**:367–373, http://dx.doi.org/10.1016/S0022-3476(80)80674-3.

133. Mortensen R, Nissen TN, Fredslund S, Rosenkrands I, Christensen JP, Andersen P, Dietrich J: **Identifying protective Streptococcus pyogenes vaccine antigens recognized by both B and T cells in human adults and children.** *Sci Rep* 2016, **6**:22030, http://dx.doi.org/10.1038/srep22030.

134. Halperin RF, Stafford P, Emery JS, Navalkar K, Johnston SA: **GuiTope: An application for mapping random-sequence peptides to protein sequences**. *BMC Bioinformatics* 2012, **13**:1, http://dx.doi.org/10.1186/1471-2105-13-1.

135. Peters B, Sidney J, Bourne P, Bui HH, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O, et al.: **The immune epitope database and analysis resource: From vision to blueprint**. *PLoS Biol* 2005, **3**:0379–0381, http://dx.doi.org/10.1371/journal.pbio.0030091.

136. Caoili SEC: **Benchmarking B-cell epitope prediction with quantitative dose-response data on antipeptide antibodies: Towards novel pharmaceutical product development**. *Biomed Res Int* 2014, **2014**, http://dx.doi.org/10.1155/2014/867905.

137. Alter-Wolf S, Blomberg BB, Riley RL: **Deviation of the B cell pathway in senescent mice is associated with reduced surrogate light chain expression and altered immature B cell generation, phenotype, and light chain expression.** *J Immunol* 2009, **182**:138–47, http://dx.doi.org/10.1016/j.bbi.2008.05.010.

138. Heikkinen T, Järvinen A: **The common cold**. *Lancet* 2003, **361**:51–59, http://dx.doi.org/10.1016/S0140-6736(03)12162-9.

139. Palermo A, Weber LK, Rentschler S, Isse A, Sedlmayr M, Herbster K, List V, Hubbuch J, Löffler FF, Nesterov-Müller A, et al.: **Identification of a tetanus toxin specific epitope in single amino acid resolution**. *Biotechnol J* 2017, **1700197**:1700197, http://dx.doi.org/10.1002/biot.201700197.

140. Bozekowski JD, Graham AJ, Daugherty PS: **High-titer antibody depletion enhances discovery of diverse serum antibody specificities**. *J Immunol Methods* 2018, **455**:1–9, http://dx.doi.org/10.1016/j.jim.2018.01.003.

141. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH: **UniRef: Comprehensive and non-redundant UniProt reference clusters**. *Bioinformatics* 2007, **23**:1282–1288, http://dx.doi.org/10.1093/bioinformatics/btm098.

142. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al.: **Biopython: Freely available Python tools for computational molecular biology and bioinformatics**. *Bioinformatics* 2009, **25**:1422–1423, http://dx.doi.org/10.1093/bioinformatics/btp163.

143. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al.: **Scikit-learn: Machine learning in Python**. *J Mach Learn Res* 2012, **12**:2825–2830, http://dx.doi.org/10.1007/s13398-

014-0173-7.2.

144. Hunter JD: **Matplotlib: A 2D graphics environment**. *Comput Sci Eng* 2007, **9**:99–104, http://dx.doi.org/10.1109/MCSE.2007.55.

145. Nybakken GE, Oliphant T, Johnson S, Burke S, Diamond MS, Fremont DH: **Structural basis of West Nile virus neutralization by a therapeutic antibody**. *Nature* 2005, **437**:764–769, http://dx.doi.org/10.1038/nature03956.

146. Ahmad TA, Eweida AE, Sheweita SA: **B-cell epitope mapping for the design of vaccines and effective diagnostics**. *Trials Vaccinol* 2016, **5**:71–83, http://dx.doi.org/10.1016/j.trivac.2016.04.003.

147. Cohen JI: **Epstein-Barr virus infection**. *N Engl J Med* 2000, **343**:481–492, http://dx.doi.org/10.1056/NEJM200008173430707.

148. Ballew JT, Murray JA, Collin P, Maki M, Kagnoff MF, Kaukinen K, Daugherty PS: **Antibody biomarker discovery through in vitro directed evolution of consensus recognition epitopes**. *Proc Natl Acad Sci* 2013, **110**:19330–19335, http://dx.doi.org/10.1073/pnas.1314792110.

149. Amornsiripanitch N, Hong S, Campa MJ, Frank MM, Gottlin EB, Patz EF: **Complement factor H autoantibodies are associated with early stage NSCLC**. *Clin Cancer Res* 2010, **16**:3226–3231, http://dx.doi.org/10.1158/1078-0432.CCR-10-0321.

150. Bushey RT, Moody MA, Nicely NL, Haynes BF, Alam SM, Keir ST, Bentley RC, Roy Choudhury K, Gottlin EB, Campa MJ, et al.: **A therapeutic antibody for cancer, derived from single human B cells**. *Cell Rep* 2016, **15**:1505–1513, http://dx.doi.org/10.1016/j.celrep.2016.04.038.

151. Burton DR, Desrosiers RC, Doms RW, Koff WC, Kwong PD, Moore JP, Nabel GJ, Sodroski J, Wilson IA, Wyatt RT: **HIV vaccine design and the neutralizing antibody problem**. *Nat Immunol* 2004, **5**:233–236, http://dx.doi.org/10.1038/ni0304-233.

152. Ferrari G, Haynes BF, Koenig S, Nordstrom JL, Margolis DM, Tomaras GD: **Envelope-specific antibodies and antibody-derived molecules for treating and curing HIV infection**. *Nat Rev Drug Discov* 2016, **15**:823–834, http://dx.doi.org/10.1038/nrd.2016.173.

153. Wang CY, Walfield AM: **Site-specific peptide vaccines for immunotherapy and immunization against chronic diseases, cancer, infectious diseases, and for veterinary applications**. *Vaccine* 2005, **23**:2049–2056, http://dx.doi.org/10.1016/j.vaccine.2005.01.007.

154. Van Regenmortel MHV: **Specificity, polyspecificity, and heterospecificity of antibody-antigen recognition**. *J Mol Recognit* 2014, **27**:627–639, http://dx.doi.org/10.1002/jmr.2394.

155. Thomsen MCF, Nielsen M: **Seq2Logo: A method for construction and**

visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res* 2012, **40**:281–287, http://dx.doi.org/10.1093/nar/gks469.

156. Gershoni JM, Roitburd-Berman A, Siman-Tov DD, Freund NT, Weiss Y: **Epitope mapping: The first step in developing epitope-based vaccines**. *BioDrugs* 2007, **21**:145–156, http://dx.doi.org/10.2165/00063030-200721030-00002.

157. Adams GP, Weiner LM: **Monoclonal antibody therapy of cancer**. *Nat Biotechnol* 2005, **23**:1147–1157, http://dx.doi.org/10.1038/nbt1137.

158. AREDS Research Group: **A randomized, placebo-controlled, clinical trial of high-dose supplementation with vitamins C and E, beta carotene, and zinc for age-related macular degeneration and vision loss: AREDS report no. 8.** *Arch Ophthalmol (Chicago, Ill 1960)* 2001, **119**:1417–1436, http://dx.doi.org/10.1016/j.bbi.2008.05.010.

159. Reeder SB, Hu HH, Sirlin CB, Group LI, Diego S: **Transcriptomic segregation of human autoantigens useful for the diagnosis of autoimmune diseases**. *Mol Diagn Ther* 2016, **36**:1011–1014, http://dx.doi.org/10.1002/jmri.23741.Proton.

160. Saeki K, Miura Y, Aki D, Kurosaki T, Yoshimura A: **The B cell-specific major raft protein, Raftlin, is necessary for the integrity of lipid raft and BCR signal transduction**. *EMBO J* 2003, **22**:3015–3026, http://dx.doi.org/10.1093/emboj/cdg293.

161. Chen R, Amoui M, Zhang Z, Mardon G: **Dachshund and eyes absent proteins form a complex and function synergistically to induce ectopic eye development in Drosophila**. *Cell* 1997, **91**:893–903, http://dx.doi.org/10.1016/S0092-8674(00)80481-X.

162. Brown JS, Hussell T, Gilliland SM, Holden DW, Paton JC, Ehrenstein MR, Walport MJ, Botto M: **The classical pathway is the dominant complement pathway required for innate immunity to Streptococcus pneumoniae infection in mice.** *Proc Natl Acad Sci U S A* 2002, **99**:16969–74, http://dx.doi.org/10.1073/pnas.012669199.

163. Zipfel PF, Lauer N, Skerka C: *Inflammation and retinal disease: Complement biology and pathology*. Springer New York; 2010.

164. Kalimo KO, Marttila RJ, Granfors K, Viljanen MK: **Solid-phase radioimmunoassay of human immunoglobulin M and immunoglobulin G antibodies against herpes simplex virus type 1 capsid, envelope, and excreted antigens.** *Infect Immun* 1977, **15**:883–9.

165. De Marchi CR, Di Noia JM, Frasch ACC, Neto VA, Almeid IC, Buscaglia CA: **Evaluation of a recombinant Trypanosoma cruzi mucin-like antigen for serodiagnosis of Chagas' disease**. *Clin Vaccine Immunol* 2011, **18**:1850–1855, http://dx.doi.org/10.1128/CVI.05289-11.

166. Lundberg IE: **Cardiac involvement in autoimmune myositis and mixed connective tissue disease**. *Lupus* 2005, **14**:708–712, http://dx.doi.org/10.1191/0961203305lu2205oa.

167. Hamaguchi Y, Mehra S, Walker J, Patterson K, Fritzler MJ: **Autoantibodies in systemic sclerosis**. *Autoimmun Rev* 2013, **12**:340–354, http://dx.doi.org/10.1007/978-4-431-55708-1_14.

168. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, Peters N, Adlem E, Tivey A, Aslett M, et al.: **Comparative genomic analysis of three Leishmania species that cause diverse human disease**. *Nat Genet* 2007, **39**:839–847, http://dx.doi.org/10.1038/ng2053.

169. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream MA, Adlem E, Aert R, et al.: **The genome of the kinetoplastid parasite, Leishmania major**. *Science (80-)* 2005, **309**:436–442, http://dx.doi.org/10.1126/science.1112680.

170. Backert L, Kohlbacher O: **Immunoinformatics and epitope prediction in the age of genomic medicine**. *Genome Med* 2015, **7**:119, http://dx.doi.org/10.1186/s13073-015-0245-0.

171. Ayoglu B, Schwenk JM, Nilsson P: **Antigen arrays for profiling autoantibody repertoires**. *Bioanalysis* 2016, **8**:1105–1126, http://dx.doi.org/10.4155/bio.16.31.

172. Ho DWT, Field PR, Sjogren-Jansson E, Jeansson S, Cunningham AL: **Indirect ELISA for the detection of HSV-2 specific IgG and IgM antibodies with glycoprotein G (gG-2)**. *J Viroloaical Methods* 1992, **36**:249–264, http://dx.doi.org/10.1016/0166-0934(92)90056-J.

173. Ahmad TA, Eweida AE, Sheweita SA: **B-cell epitope mapping for the design of vaccines and effective diagnostics**. *Trials Vaccinol* 2016, **5**:71–83, http://dx.doi.org/10.1016/j.trivac.2016.04.003.

174. Niespodziana K, Napora K, Cabauatan C, Focke-Tejkl M, Keller W, Niederberger V, Tsolia M, Christodoulou I, Papadopoulos NG, Valenta R: **Misdirected antibody responses against an N-terminal epitope on human rhinovirus VP1 as explanation for recurrent RV infections**. *FASEB J* 2012, **26**:1001–1008, http://dx.doi.org/10.1096/fj.11-193557.